electronics

MDPI

Article

EDANet: Efficient Dynamic Alignment of Small Target Detection Algorithm

Gaofeng Zhu [1], Fenghua Zhu [2], Zhixue Wang [1,*], Shengli Yang [3] and Zheng Li [1]

[1] School of Rail Transit, Shandong Jiaotong University, Jinan 250300, China
[2] Chinese Academy of Sciences, Beijing 100190, China
[3] CHN Energy Investment Group Co., Ltd., Beijing 100190, China
* Correspondence: gaofeng_zhu0020@163.com

Abstract: Unmanned aerial vehicles (UAVs) integrated with computer vision technology have emerged as an effective method for information acquisition in various applications. However, due to the small proportion of target pixels and susceptibility to background interference in multi-angle UAV imaging, missed detections and false results frequently occur. To address this issue, a small target detection algorithm, EDANet, is proposed based on YOLOv8. First, the backbone network is replaced by EfficientNet, which can dynamically explore the network size and the image resolution using a scaling factor. Second, the EC2f feature extraction module is designed to achieve unique coding in different directions through parallel branches. The position information is effectively embedded in the channel attention to enhance the spatial representation ability of features. To mitigate the low utilization of small target pixels, we introduce the DTADH detection module, which facilitates feature fusion via a feature-sharing interactive network. Simultaneously, a task alignment predictor assigns classification and localization tasks. In this way, not only is feature utilization optimized, but also the number of parameters is reduced. Finally, leveraging logic and feature knowledge distillation, we employ binary probability mapping of soft labels and a soft label weighting strategy to enhance the algorithm's learning capabilities in target classification and localization. Experimental validation on the UAV aerial dataset VisDrone2019 demonstrates that EDANet outperforms existing methods, reducing GFLOPs by 39.3% and improving Map by 4.6%.

Keywords: deep learning; object detection; attention mechanism; computer vision; smart transportation

Academic Editor: Felipe Jiménez

Received: 5 December 2024
Revised: 3 January 2025
Accepted: 5 January 2025
Published: 8 January 2025

Citation: Zhu, G.; Zhu, F.; Wang, Z.; Yang, S.; Li, Z. EDANet: Efficient Dynamic Alignment of Small Target Detection Algorithm. *Electronics* **2025**, *14*, 242. https://doi.org/10.3390/electronics14020242

1. Introduction

In recent years, the rapid advancements in UAV technology, coupled with the widespread adoption of information and intelligence technologies, have significantly expanded the applications and development of UAV-based visual perception systems. Leveraging the high-altitude capabilities of UAVs in conjunction with computer vision-based target detection technology enables efficient identification and localization of targets over wide areas. These technologies are applied across various real-world scenarios, including smart logistics, ecological and forest protection, battlefield intelligence acquisition, and traffic monitoring [1,2]. Despite the advantages provided by UAVs' high altitude, multi-angles, and wide field of view, these characteristics also present significant challenges, particularly in the detection of small targets. In UAV visual perception, small targets often exhibit challenges such as low pixel resolution, complex backgrounds, and the difficulty of extract-

ing distinctive features [3,4]. Consequently, small target detection has garnered significant attention as a crucial technology for enhancing UAVs' visual perception capabilities.

In target detection, the definition of small targets is established based on two factors. First, research indicates that the smallest image size the human eye can effectively recognize is 32 × 32 pixels. Second, in convolutional neural networks, after several pooling layers, the pixel size corresponding to the final feature vector is approximately 32 × 32 pixels. Therefore, 32 × 32 pixels is considered a reasonable threshold for distinguishing small targets from regular ones.

Most small object detection applications focus on the challenge of efficiently utilizing underlying features. By continuously optimizing the parameters for feature extraction and the forward propagation of convolution kernels, the model can extract meaningful features from the data and apply these insights to predict new data, thus achieving efficient classification or detection. While typical one-stage algorithms, such as SSD [5] (Single Shot MultiBox Detector), offer fast processing speeds and are suitable for real-time applications, their detection results are often affected by background interference. YOLO (You Only Look Once), an efficient single-stage detection method, can directly predict both the category and location of targets, making it widely used in real-time detection tasks. However, these methods still encounter limitations in detecting small targets. Two-stage algorithms, such as Faster R-CNN [6] (Faster Region-based Convolutional Neural Network) and Mask R-CNN (Mask Region-based Convolutional Neural Network), generate candidate regions through a region proposal network and perform classification and bounding box regression on these regions. While these algorithms achieve higher accuracy, they have slower inference speeds and higher computational overhead. Transformer-based detection algorithms, such as DETR [7] (End-to-End Object Detection with Transformers), improve the model's ability to capture spatial relationships through a self-attention mechanism and optimize the matching process between predicted and real frames using a bipartite graph matching algorithm. While detection accuracy is improved, these methods increase computational overhead, resulting in longer training times and slower inference speeds, particularly for small target detection. Most algorithm development is based on conventional target design; however, there are certain limitations when applied to small target detection scenarios. These limitations can be summarized as the following challenges through inductive analysis:

1. Due to the limitations of the imaging angle and adverse conditions such as low light or cloudy weather, confusion between the background and target is common. Furthermore, elements such as clouds, ground debris, and buildings in the background share similar characteristics with the target, making small targets susceptible to interference. Effectively distinguishing target features from background features in complex, dynamic environments and extracting small target features with high recognition accuracy has become a core challenge in improving detection performance.
2. Each image typically contains multiple small targets, and the pixel ratio of these targets is relatively low. As a result, in the high-dimensional feature space, the representation of small targets is sparse, which increases the difficulty of extraction and recognition. This sparsity often leads to false detections or missed detections, particularly in complex scenes, significantly affecting the accuracy of small target detection.
3. The low resolution of small targets not only hampers the accurate extraction of local features but also reduces the efficiency of global feature utilization. The performance of small target detection is often constrained by both resolution and feature extraction capabilities. Optimizing these factors will directly enhance the effectiveness of the detection algorithm.

2. Related Work

To comprehensively enhance the utilization efficiency of underlying detailed features in small target detection tasks, researchers have conducted extensive investigations from multiple perspectives. Some scholars achieve efficient feature extraction by replacing the backbone network with a more efficient alternative. For example, the method described in [8] substitutes the ShuffleNet-v1 backbone with a combination of group convolution and channel rearrangement techniques, enabling information exchange between groups and thereby improving the model's expressive capability. The method described in [9] optimizes the network structure by replacing the MobileNet-v3 backbone and incorporates deep convolution alongside automated architecture search techniques to develop a more efficient feature extraction backbone, thereby improving the utilization of underlying detailed information.

At the same time, the application of attention mechanisms has also enhanced the precise localization of small targets. The approach reported in [10] further improved the effective utilization of information within the input feature space and channels by dividing the input features along the channel dimension, separately extracting spatial and channel attention features, and then adaptively weighting them to form a collaborative attention mechanism. In small target detection tasks, the method described in [11] integrates the SimAM (Simple, Parameter-Free Attention Module) attention mechanism, which eliminates the need for additional parameters. It interacts with channel information to construct three-dimensional attention weights, thereby enhancing the model's classification ability and significantly improving the recognition accuracy of both similar inter-class and intra-class instances.

To address the impact of complex backgrounds on small target detection, the method described in [12] focuses on aggregating target areas at each scale and learns the proposal boxes and corresponding features of these areas through a scale-separated, learnable proposal mechanism. This mechanism allows the number of proposal boxes to be dynamically adjusted at each scale, thereby enabling accurate identification of targets across different scales and distributions with low computational cost. The method described in [3] introduces a spatial information blending module that effectively mitigates the issue of target information blur caused by object occlusion. By merging feature information through multi-branch atrous convolution, the network enhances its receptive field and feature reuse capabilities, facilitating information interaction and obtaining more accurate spatial position data, which enables the model to learn detailed features across various dimensions. The method described in [13] constructs a multi-head self-attention mechanism for cross-spatial learning, assigns weights to different features based on their correlation, establishes long-range contextual relationships of small object scale information, and extracts critical spatial details. The method described in [14] utilizes an adaptive threshold focus loss function to decouple the target from the background, adjusting the loss weight via an adaptive mechanism to compel the model to focus more on target features. Additionally, normalizing the Gaussian Wasserstein distance mitigates convergence difficulties caused by the extreme sensitivity of bounding box regression to small targets in infrared imagery.

Regarding the feature fusion module, the method described in [15] employs feature-enhanced spatial pyramid pooling and cross-stage partial connections to effectively mitigate feature loss during the extraction stage and enhance the feature representation capabilities for small targets. Additionally, it addresses feature loss by integrating an attention-guided Max pooling module. The method described in [16] employs an adaptive fusion strategy alongside a feature pyramid network enriched with semantic information to effectively resolve potential information conflicts between layers during the fusion process. The method described in [17] introduced a multiplexed adaptive spatial pyramid pooling module that

mitigates the impact of complex background clutter through multiplexed pooling operations and adaptive fusion. This module also includes a novel feature pyramid network with an adaptive downsampling module, which reduces information loss from downsampling and enhances the capability for multi-scale small object detection.

With regard to optimizing the object detection head, the method described in [18] achieves more flexible and efficient feature extraction by incorporating pixel encoders and decoders that exchange spatial and channel dimensions within the head layer. The method described in [19] combined a focused attention mechanism to amplify important feature expressions while suppressing irrelevant ones, thereby enhancing the network's capability to represent target features and ensuring the retention and enhancement of crucial features. The method described in [20] designed the Swin Transformer Prediction Head, a convolutional prediction head that introduces an advanced self-attention mechanism and reduces computational complexity through a moving window design, all while maximizing the preservation of feature information. The method described in [21] introduces an additional small target prediction head, replacing the large target prediction head to achieve higher accuracy in small target detection. By performing convolution operations at different scales across multiple branches, rich, multi-scale feature information can be extracted, thereby enhancing the network's ability to capture fine-grained details in small objects.

Since its introduction by [22] in 2006, knowledge distillation has evolved into a key method for model compression. The method described in [23] employed knowledge distillation for neural network compression, guiding the learning of the student network using the soft labels from the teacher network. Recently, knowledge distillation has been widely applied in image classification tasks and has progressively extended to object detection. The method described in [24] utilizes the intermediate layer features of the teacher network to guide the student network's learning, significantly enhancing the performance of smaller models. Knowledge distillation offers an effective approach to enhance small object detection performance without extensive modifications to the model structure. The method described in [25] introduced a dual masking knowledge distillation method to capture spatial and channel information cues, which significantly improved feature reconstruction capabilities.

In summary, although different aspects of optimization methods have achieved certain improvements in the field of object detection, the effect of small object detection is limited, especially in the face of background interference, and low pixel performance still needs to be further improved. Therefore, based on the YOLOv8n algorithm, this paper focuses on solving the challenges in the process of small object detection, and proposes an improved algorithm.

3. Improvements to the Model Architecture

The network structure of YOLOv8n is very simple. It can be divided into three main parts: a backbone for feature extraction, a neck for feature fusion, and a head for target category and location regression detection.

Based on the framework of the basic network, the new algorithm is designed and optimized, and the network structure is shown in Figure 1. Aiming at the challenge of small target detection in aerial image datasets, targeted improvements are mainly made from the following aspects.

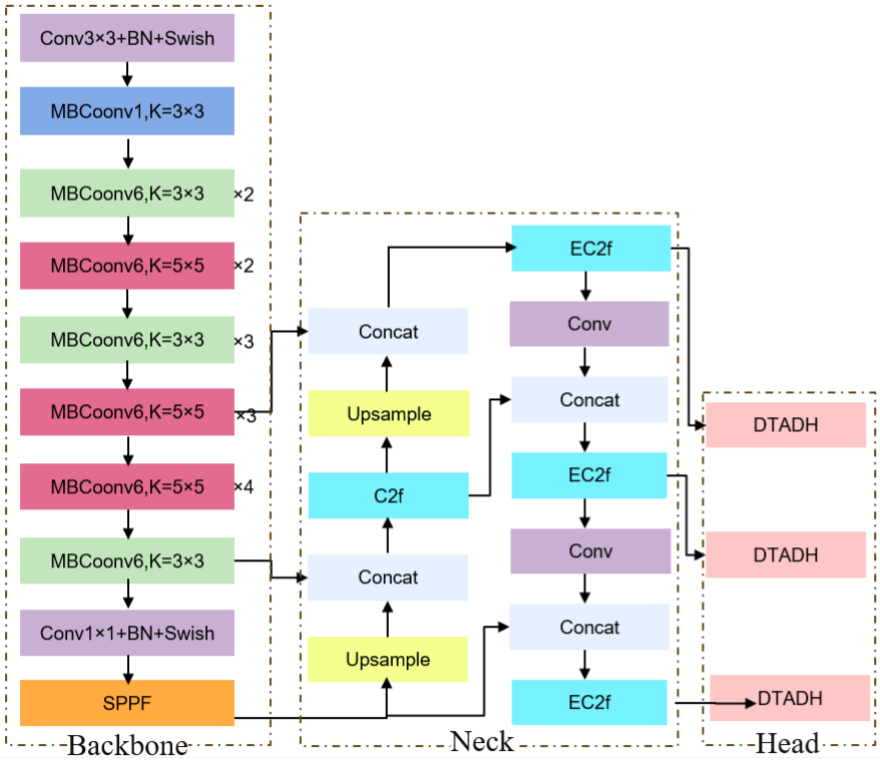

Figure 1. Overall structure of the optimization algorithm.

In response to the significant challenge of extracting and locating small targets, this paper proposes a general small target detection algorithm EDANet. The research focuses on four key aspects: feature extraction, feature fusion, a shared detection head, and knowledge distillation. The goal is to address these challenges and enhance the overall performance of small target detection. The primary contributions of this work are as follows:

1. Scalable and Efficient Detection Backbone: The CSPDarknet (Cross-Stage Partial Darknet) network structure in the original backbone is stable; however, during small target detection, key features are often lost as the convolutional layers deepen. To address this issue, the original backbone network is replaced by EfficientNet [26], which introduces a composite scaling factor to adjust the depth, width, and resolution of the backbone network, ensuring adaptability to various application tasks and computational resource constraints. Furthermore, the approach leverages substantial data and computational resources to learn features directly from raw data, thereby eliminating the complexity of manual feature engineering. This adaptation better aligns with the target distribution and task requirements in small target datasets, significantly enhancing the algorithm's detection performance.
2. EC2f Extraction Module: The C2f (Cross-Stage Feature Fusion) module in the basic algorithm performs feature extraction at different scales through continuous convolution, but this is not conducive to the detailed feature extraction of small targets. To address this limitation, the EMA (Efficient Multi-Scale Attention) mechanism and a residual structure are introduced, forming the cross-space multi-dimensional feature extraction module (EC2f). This module effectively extracts fine-grained features of small targets. The parallel branch design of the attention mechanism encodes features in multiple directions, embeds positional information into channel attention, and enhances spatial representation capabilities. This design ensures the effective extraction of fine-grained features from small targets. Finally, feature fusion within the residual structure integrates information from different levels and stages, enabling the capture of richer features within the image.

3. Dynamic Task Alignment Detection Head (DTADH): The proposed algorithm incorporates a multi-level shared feature module, constructed using a parallel decoupling head combined with Group Normalization (Group-Norm) and a residual network, to model feature interactions effectively. The task alignment predictor assigns target classification and localization tasks, optimizing feature utilization and reducing the parameter count. Additionally, the attention mechanism is integrated to emphasize key features for classification and localization tasks, thereby enhancing the detection head's ability to identify small targets.
4. Hybrid Distillation Learning Module: Building on the DTADH model design and leveraging the strengths of logical distillation, a parallel distillation loss function is constructed for positioning and classification tasks. Simultaneously, a feature map-based distillation loss function is developed to address feature distillation, resulting in a hybrid distillation learning module through weighted summation. This module integrates the advantages of logic and feature knowledge distillation, enabling end-to-end training. By transferring rich features and logical inputs, the algorithm's detection performance is further enhanced, effectively addressing the challenges of positioning accuracy and insufficient feature extraction in small target detection.

3.1. Efficient Feature Extraction Backbone Network

When traditional convolutional neural network (CNN) algorithms seek to enhance object detection accuracy, they typically do so by modifying the network's depth, width, or the resolution of input images. For example, the VGG network improves performance by increasing layer depth and convolutional complexity, while DenseNet and ResNet focus on adjusting network width to achieve high detection accuracy. However, augmenting depth and width can lead to challenges such as vanishing or exploding gradients.

EfficientNet addresses these challenges through compound scaling, which balances the depth, width, and image resolution of the network. It effectively prevents gradient issues by incorporating the Swish activation function and residual branches. Additionally, EfficientNet employs Dropout technology to randomly deactivate neurons, which helps prevent overfitting, and carefully manages parameter calculations to improve accuracy. Following neural architecture search, EfficientNet constructs the Efficient-B0 network, as detailed in Table 1. The network's overall operation can be summarized as follows:

$$N(d, w, r) = \underset{i=1,\ldots s}{e} F_i^{L_i}(X_{[H_i, W_i, C_i]}) \tag{1}$$

here, $\underset{i=1,\ldots s}{e}$ represents a multiplication operation; $F_i^{L_i}$ means that the F_i operation is repeated L_i times in Stage i; S is the characteristic matrix of the input stage i; and $[H_i, W_i, C_i]$ is the height, width, and number of channels of the input.

Table 1. The Efficient-B0 network structure table.

Stage (i)	Operate (Fi)	Resolution (Hi × Wi)	Channels (Ci)	Layers (Li)
1	Conv3 × 3	224 × 224	32	1
2	MBConv1, k = 3 × 3	112 × 112	16	1
3	MBConv6, k = 3 × 3	112 × 112	24	2
4	MBConv6, k = 5 × 5	56 × 56	40	2
5	MBConv6, k = 3 × 3	28 × 28	80	3
6	MBConv6, k = 5 × 5	14 × 14	112	3
7	MBConv1, k = 5 × 5	14 × 14	192	4
8	MBConv1, k = 3 × 3	7 × 7	320	1
9	Conv1 × 1 and Pooling and FC	7 × 7	1280	1

EfficientNet utilizes the MBConv module for backbone construction, employing an inverted residual structure combined with carefully designed layers to enhance the algorithm's representational capacity. The 1 × 1 convolution increases the dimension of the input features, while the k × k DepthwiseConv2D convolution extracts spatial features. Additionally, the SE (Squeeze and Excitation) structure interactively models features across different channels, and the dimension of the modeled feature map is reduced by another 1 × 1 convolution. Dropout is also employed to mitigate the risk of overfitting. This inverted residual design philosophy allows MBConv to provide a strong representational capability while maintaining computational efficiency, making it well suited for efficient low-level feature extraction in small object detection networks.

By applying this composite scaling method, the network adapts effectively to the small target detection environment. The integration of the SE module ensures that key features of small targets are retained and highlighted, allowing for more efficient extraction of fine-grained features. This approach not only improves detection accuracy and reliability but also enhances the robustness of the algorithm in complex environments, making small object detection more effective and practical across various real-world scenarios.

3.2. Feature Extraction Module

In this paper, the EC2f module is designed to address the challenges of effective feature extraction in scenarios with dense object arrangements and complex backgrounds in small object detection tasks. This module leverages the multi-level interaction features of the C2f module and integrates the multi-scale feature extraction capabilities of the EMA module. By comprehensively considering both local and global features, the EC2f module enhances the ability to distinguish adjacent targets in densely packed scenarios and extracts useful information amidst complex backgrounds, thereby enabling the efficient capture of small target detail features.

The network structure of the EC2f feature extraction module is depicted in Figure 2. This module captures cross-dimensional feature information through a combination of a triple parallel branch structure and coordinate attention. By interactively modeling vertical and horizontal channel information, the vertical direction enhances spatial dependency across different depths, while the horizontal direction captures target location information within the channels. This location information is integrated into channel attention, improving the spatial representation of features. The design of the EC2f module aims for efficient feature extraction while maintaining the advantages of minimal depth and low latency. Consequently, it effectively extracts and utilizes feature information in scenarios with densely arranged targets and complex backgrounds, thereby enhancing the performance of small target detection.

Specifically, without reducing the channel dimension, the given $Input \in R^{C\times H\times W}$ is divided into G sub-features $X = [X_0, X_1, \ldots X_i \ldots X_{G-1}]G = C$ according to the direction of the dimension. The low-level features of small objects are grouped to learn different semantic information and strengthen the expression of sub-features in the feature map. At the same time, the local receptive field collects multi-scale spatial information to generate high-resolution feature maps. The parallel branch uses global average pooling to encode the weights of different directions and convert them into the corresponding dimensional shape $R^{1\times C//G} \times R^{C//G\times HW}$ for cross-channel information exchange, and effectively adjusts the weights of parameters between different channels. The formula is as follows:

$$Z_c^w = \frac{1}{H} \sum_{0\le j<H} x_c(j, w) \tag{2}$$

$$Z_c = \frac{1}{W} \sum_{0 \le j < W} x_c(h, i) \tag{3}$$

$$Z_c = \frac{1}{H \times W} \sum_j^H \sum_i^W x_c(i, j) \tag{4}$$

where x_c represents the input feature of the c channel; Z_c^w represents the output of the channel width of w, applied in the vertical direction; and Z_c^h represents the output of the c channel width of h, applied in the horizontal direction.

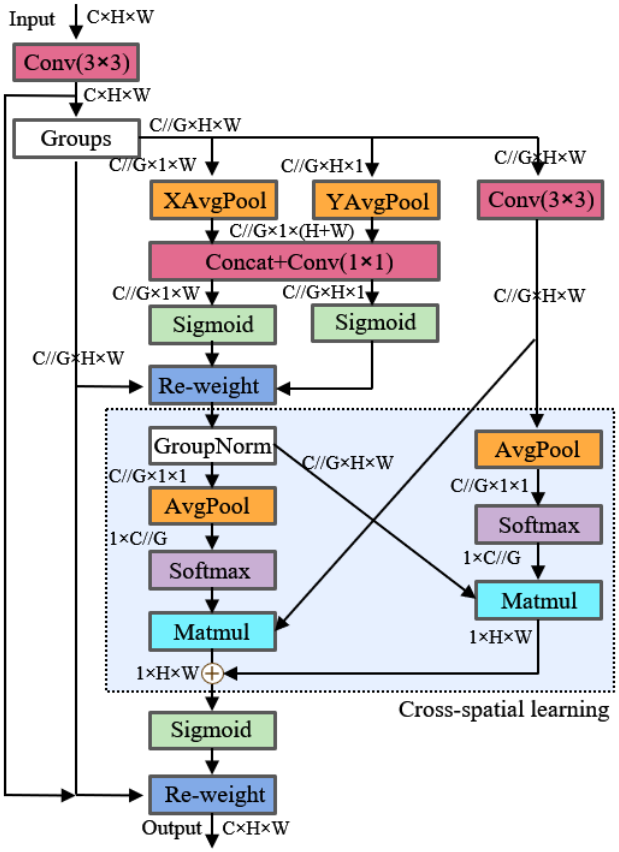

Figure 2. The network structure of the EC2f feature extraction module. Here, Groups means that the feature map is divided into G sub-features in the direction of the dimension.

Two-dimensional global average pooling and the Softmax nonlinear activation function are applied to process the feature maps in both directions. The minimum branch output dimension can be adjusted to meet specific requirements while preserving the precise spatial structure information of the feature channels. This approach extends the dimension transformation capability of the feature space, efficiently extracts dependencies among the three channels, and retains spatial structure information in the channel dimension, thereby reducing computational overhead.

The attention mechanism is integrated to model information across different dimensions, preserving both accurate location and semantic information. Contextual information from feature maps at various scales is fused to generate granular attention. A three-branch parallel network structure is designed to efficiently handle dependencies between different depth information through cross-dimensional feature interaction, simultaneously capturing the underlying details of small targets across varying receptive fields. The output of the EC2f module is crucial for channel attention and the interaction of detailed information. It excels at distinguishing the features of small targets at different scales, effectively mitigating background interference and capturing critical information.

3.3. Dynamic Task Alignment Detection Head

The detection head in the proposed algorithm adopts a decoupled head structure, dividing the category classification and bounding box regression tasks into two parallel networks for simultaneous processing. This approach eliminates mutual interference between tasks and optimizes the efficiency of each task execution. Task-specific streamlined networks are designed to reduce the parameter count and computational complexity, thereby enhancing the algorithm's generalization and robustness. However, the parallel structure lacks sufficient feature interaction, leading to suboptimal feature utilization. To address this, we propose a dynamic alignment decoupling head that enhances feature interaction while maintaining the decoupled structure.

The basic algorithm's detection head uses parallel and independent classification and localization networks, which lack sufficient feature interaction between task branches, affecting the detection and localization of small targets. To overcome this, an optimized feature-sharing network is designed, incorporating the alignment predictor [27] to improve the classification and localization of small targets.

As shown in Figure 3, a multi-level shared feature module is constructed by integrating the GroupNorm [28] module with a residual structure. This module processes input feature maps of different scales through convolution, calculates corresponding global weights after normalization, and utilizes the attention mechanism to dynamically decompose and calculate task features. This design enables interactive perception of classification and localization tasks, enhancing the detection head's performance in recognizing and detecting small targets. The shared convolution approach significantly reduces the number of parameters and dependence on batch size, mitigating the risk of overfitting and making the algorithm more lightweight.

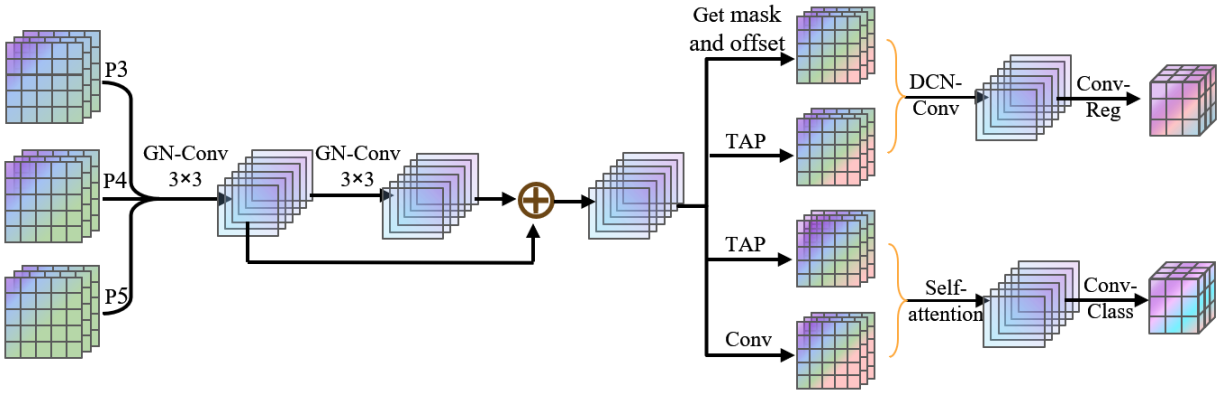

Figure 3. Dynamic task alignment detection head network structure diagram. The left side of the image represents the construction of a feature-sharing network; the right side represents the task alignment predictor for dynamic task allocation to achieve parallel classification and positioning functions. Here, P3, P4, and P5 represent feature maps under different convolution depths.

For shared convolutional output features, dynamic task allocation is managed by the task alignment predictor, which learns interactive task features from multiple convolutional layers to generate joint features. In the positioning branch, DCNv2 (Deformable Convolutional Networks Version 2) and 3 × 3 mask convolution are employed to generate dynamic offsets and masks for the joint features after feature fusion. This enables the definition of local attention areas, allowing the model to focus more on regions likely containing small objects. The dynamic offset adjusts the shape and position of the convolution kernel to accommodate geometric changes in the input feature map, thereby enhancing adaptability

to target shape variations and enabling dynamic feature selection of interactive features. The specific operation process is described by the following formula:

$$w = s(f_{Conv2}(s(f_{Conv1}(X^{inter})))) \tag{5}$$

$$X_k^{task} = w_k \times X_k^{inter}, k \in (1, 2, \ldots N) \tag{6}$$

$$Z_k^{task} = f_{c2}(s(f_{c1}(X_k^{task}))) \tag{7}$$

$$y(p) = \sum_{k=1}^{k} w(p_n) \cdot Z_k^{task}(p + p_n + \Delta p_n) \cdot \Delta m_k \tag{8}$$

where w is calculated from the cross-layer task interaction features, capturing dependencies between layers; X^{inter} represents the interaction feature of the input; w_k represents the dependency between the K element of the interactive feature and different feature layers; Xk^{task} is the splicing feature of the interactive feature layer and Z_k^{task} corresponds to the key feature derived by the feature extractor; $x(p)$ represents the mapping of input features; $y(p)$ represents the feature at position P on the output feature map; $w(pn)$ is the weight of the corresponding offset position; Δpn is the learned offset; $y(p)$ denotes the learned addition scalars; and P is the center of the convolution kernel and p_n specifies the offset.

Finally, to address the issue of inconsistent target scales detected by each detection head, a scale layer is employed to adjust the features and output the final detection results. By leveraging the shared feature module and combining it with the task alignment predictor for feature allocation, the task alignment structure enhances the accuracy of selecting anchor points that match small targets. This optimization improves the sample allocation strategy and increases the detection rate of small targets. Additionally, this approach simplifies the detection process and effectively reduces the computational burden of the module, especially on resource-constrained devices.

3.4. Hybrid Knowledge Distillation Module

Knowledge distillation is a model compression technique that imitates a large trained teacher model by training a small student model. This approach can not only transfer the knowledge of the teacher model, but also enable the student model to achieve better performance while maintaining a small computational cost. Since the teacher model has been trained on the image, based on the transfer of the prior knowledge of the teacher model, the student model can combine the prior knowledge to extract the features of more small targets and locate the position of more small target objects during the training process.

The core idea of knowledge distillation is that the teacher model not only learns how to accurately classify or predict, but also learns many implicit patterns and associations during the training process, and this knowledge is reflected in the intermediate layer representation of the model. With an appropriate loss function, the student model can learn to mimic these intermediate layer features of the teacher model, not just the final output. The advantage of this approach is that the student model can inherit the generalization ability of the teacher model, while it is more suitable for deployment on edge devices or mobile devices due to its smaller number of parameters and faster inference speed. Therefore, in this paper, we incorporate the EC2f and DTADH modules into the YOLOv8l model to obtain the final experimental results as the teacher model.

3.4.1. Construction of Logical Distillation Loss Function

Currently, there are two main loss functions used to implement knowledge distillation: logical loss and feature loss. The logic loss focuses on the difference between soft targets at the output layer, while the feature loss measures the difference between the student model and the teacher model at the feature representation level. In this paper, by combining

these two losses and performing weighted hybrid control, the student model can learn the knowledge of the teacher model at different levels. The specific structure is shown in Figure 4.

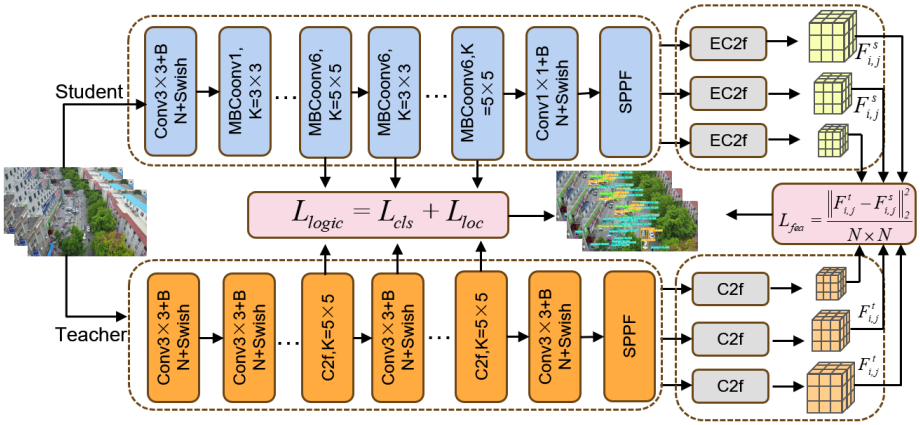

Figure 4. Schematic diagram of the structure of mixed knowledge. The left side of the image represents the calculation method of the Logit loss function; the right side represents the feature loss function based on the feature map; the final weighted calculation is performed to obtain the final hybrid distillation loss function.

According to the data analysis of the VisDrone aerial dataset, the imaging of aerial images has the problems of uneven target distribution and serious background interference. In the common logic distillation process, when the target class probability mapping is carried out on the teacher model, the background weight is increased, which affects the class probability mapping of small targets. Therefore, the sigmoid function is used to perform soft label binary probability mapping on the logic input, and the classification probability and cross-entropy loss between the labels are calculated. The weight factor is introduced for the weighting strategy of soft labels to highlight the importance of key samples. At the same time, the output bounding boxes of the teacher model and the student model are used for logic distillation, and the positioning expression of the predicted target box in the corresponding category target is given. box_i^s and box_i^t calculate the localization loss between the two, and then stimulate the student model to learn the logic output of the professor model, effectively transferring the object localization knowledge from the professor to the student model. The specific expression is shown in the following formula.

$$p_i^t = f_{sig}^t(X), \quad p_i^s = f_{sig}^s(X) \tag{9}$$

$$L_{bce}\left(p_{i,j}^t, p_{i,j}^s\right) = -\left(\left(1 - p_{i,j}^t\right) + p_{i,j}^t \times \log\left(p_{i,j}^s\right)\right) \tag{10}$$

$$L_{cls}(X) = \sum_{i=1}^{W} \sum_{j=1}^{H} w_{i,j} \cdot L_{bce}\left(p_{i,j}^s, p_{i,j}^t\right) \tag{11}$$

where $w_{i,j}$ is the weight coefficient, which is determined by the soft label probability difference between the student model and the teacher model. $p_{i,j}^s$, $p_{i,j}^t$ represents the probability mapping of the j category in the position i in the student model and the teacher model, respectively, and box_i^s and box_i^t represent the positional expression of the prediction box in the student and teacher models.

$$Iou = \frac{box_i^s \bigcup box_i^s}{box_i^s \bigcap box_i^s}, \; L_{loc} = 1 - Iou \tag{12}$$

$$L_{\log ic} = L_{cls} + L_{loc} \tag{13}$$

3.4.2. Construction of Hybrid Distillation Loss Function

In knowledge distillation of feature maps, the goal is to transfer the rich feature knowledge learned by the teacher model to the student model, so that the student model can achieve better performance and generalization ability while maintaining a smaller model size. This is achieve by flattening the high-dimensional feature tensors of the student model and the teacher model into two-dimensional tensors. By calculating the pairwise correlation between the two channels, the corresponding high-order correlation matrix is formed. Then, the n-order difference matrix of the correlation matrix between the student model and the teacher model can be calculated through the L2 loss function, and the corresponding feature distillation loss is obtained. Feature distillation loss is a measure of the difference between the features of the student model and the features of the teacher model. By minimizing this loss, the student model can learn the key feature representations of the teacher model. The specific expression is shown in the following formula.

$$F_{i,j}^{t} = f_{normalize}^{t}\left(\left(F_i^T \cdot F_j^T\right), \dim = 2\right) \tag{14}$$

$$F_{i,j}^{s} = f_{normalize}^{s}\left(\left(F_i^S \cdot F_j^S\right), \dim = 2\right) \tag{15}$$

$$L_{fea} = \frac{\left\| F_{i,j}^{t} - F_{i,j}^{s} \right\|_2^2}{N \cdot N} \tag{16}$$

where $F_{i,j}^{t}$ represents the proof of correlation between channel i and j in the teacher model, obtained from the inner product of the feature matrix; N is the dimension size of the expanded matrix. The final distillation loss function is as follows:

$$L_{kd} = \beta_1 L_{\log ic} + \beta_2 L_{fea} \tag{17}$$

where $L_{\log ic}$ is the logical loss function; L_{fea} is the characteristic loss function; β_1 represents the weight parameter of the logical loss function; and β_2 is the weight parameter of the characteristic loss function.

Small objects account for a small proportion in the image, which leads to the imbalance of its class distribution. Effective mitigation is achieved by knowledge distillation techniques that combine logic and features. This technique is especially suitable for the case where small objects account for a small proportion in the image, resulting in extremely unbalanced positive and negative samples. The outputs of the teacher model are used as soft targets, and the binary probability mapping of these soft targets is carried out by introducing the sigmoid function to calculate the classification probability and the cross-entropy loss between labels. In addition, the output bounding boxes of the teacher model and the student model are also used for logical distillation, and by imitating the performance of the teacher model in object localization, the student model is able to learn more accurate object detection boxes, which leads to better performance in the small object detection task. Thus, the sensitivity of the model to the minority categories is improved, the model complexity and computational cost are reduced while maintaining high performance, and the generalization ability and robustness of the model are improved. In this paper, we will train the YOLOv8l model, simultaneously apply the EC2f and DTADH modules, and obtain the final experimental results, which will serve as the teacher model.

In summary, the knowledge distillation technology combining logic and features, through the binary probability mapping of soft labels, the weighting strategy of soft labels, and the feature distillation of the output of the teacher model and the student model, can effectively alleviate the class imbalance problem in small object detection, improve

the learning ability of the student model in object localization, and finally improve the performance of small object detection.

4. Experimental Platform and Dataset

4.1. Evaluation Indicators and Experimental Platform

The evaluation metrics used in this experiment include precision (P), recall (R), mean average precision (Map), the number of parameters (Params), and giga floating-point operations per second (GFLOPs). Precision evaluates the proportion of predicted positive targets that are correctly identified, reflecting the model's susceptibility to false positives. Recall assesses the proportion of actual positive targets that are successfully detected, indicating the model's capability to minimize missed detections. Map is a critical metric for assessing the overall performance of object detection models, balancing both precision and recall. In the context of small object detection, Map effectively evaluates the model's detection performance across various scales. The number of parameters indicates the model's complexity and storage requirements, while GFLOPs measure its computational complexity, which is directly related to inference speed. Reducing computational complexity can enhance inference efficiency and decrease response time, thereby improving the model's practicality and suitability for real-world applications. The formula is as follows:

$$Precision = \frac{TP}{TP + FP} \tag{18}$$

$$Recall = \frac{TP}{TP + FN} \tag{19}$$

$$Map = \frac{\sum AP_c}{num_c} \tag{20}$$

here, TP means that the instance is a positive class and the prediction is a positive class; FP means that the instance is a negative class and the prediction is a positive class; FN means that the instance is a positive class and the prediction is a negative class; AP_c means the average accuracy of each class, and num_c means the number of all categories in all datasets.

This experiment uses the framework of Pytorch, uses a dataset with a resolution of 640 × 640 for training, and the pre-training algorithm is YOLOv8n. The environment of the experiment and the training parameters are shown in Table 2. All the experimental algorithms are trained, validated, and tested under the same hyperparameters.

Table 2. List of experimental environments.

Parameter	Configuration
CPU	Intel(R)Core(TM)i5-12400F
GPU	NVIDIA GeForce RTX 3060
GPU memory	12G
operating system	Windows10
Deep learning framework	torch2.0, cuda11.6

4.2. Introduction to Experimental Datasets

The VisDrone2019 dataset was collected by the AISKYEYE team, Machine Learning and Data Mining Laboratory, Tianjin University, Tianjin City, China and mainly contains 10 small object categories [29]. The benchmark dataset includes 288 video clips consisting of 261,908 frames and 10,209 still images captured by various drone cameras covering a wide range of scenes, including photos of 14 different cities, environments, objects, and densities

in locations separated by thousands of kilometers. There are the following two difficulties in object detection in the VisDrone 2019 dataset:

1. Limited Feature Extraction: The proportion of small target pixels in the sample image is small, and the information they contain is limited, making it challenging to extract effective features;
2. Scale Variation and Interference: The scale of small targets can vary significantly in multi-angle shots, influenced by complex backgrounds, occlusion, and mutual interference between targets.

5. Experimental Results and Comparative Analysis

5.1. Comparison Experiment of Small Target Detection Effect

In order to verify that the algorithm proposed in this paper has a better effect in small object detection, it is compared with other object detection algorithms on the VisDrone2019 dataset, and their detection effects are compared. The specific results are shown in Table 3.

Table 3. Comparative experimental data sheet of different mainstream algorithms. Map represents the average detection accuracy of the target detection model, where the numeric value represents the Map value of the corresponding target category.

Models	Object Category										Map (%)
	Ped.	Ppl	Bic	Car	Van	Truck	Tri	Aw-Tri	Bus	Motor	
Faster R-CNN	20.9	14.8	7.3	51.0	29.7	19.5	14.0	8.8	30.5	21.2	21.8
CDNet	35.6	19.2	13.8	55.8	42.1	38.2	33.0	25.4	49.5	29.3	34.2
DBAINet	36.7	12.8	14.7	47.4	38.0	41.4	23.4	16.9	31.9	16.6	28.0
YOLOv4	24.8	12.6	8.6	64.3	22.4	22.7	11.4	7.6	44.3	21.7	30.7
YOLOv6s	30.7	24.4	4.24	73.4	35.1	25.3	17.8	9.76	42.2	31.7	29.5
YOLOv8n	35.9	29.5	8.85	76.3	39.9	32.4	24.5	11.8	47.7	37.7	34.5
YOLOv10n	37.8	31.2	11.1	76.9	38.0	31.8	23.2	12.9	49.0	40.7	35.3
EDANet	39.7	31.2	10.2	78.1	42.7	34.5	27.9	14.5	49.7	42.0	39.1

Comparison of the results indicates that two-stage algorithms, such as Faster R-CNN, exhibit lower Map values, suggesting they are less suitable for small target detection. In contrast, YOLO series algorithms perform well, with improvements observed in various detection categories. Specifically, the optimized YOLO algorithm shows a 4.6% increase in Map compared to the basic algorithm. It demonstrates strong detection performance for categories with small proportions, such as Pedestrian, People, and Car, and also achieves stable detection results for the Truck and Van categories, which have high similarity. Overall, the EDANet consistently outperforms other algorithms in terms of Map.

5.2. Ablation Experiment

To verify the effectiveness of EDANet, ablation experiments were conducted using the VisDrone2019 dataset. These experiments assessed the impact of different modules on the performance of the small target detection algorithm under consistent experimental conditions. YOLOv8n was chosen as the reference algorithm, with the input image resolution set to 640 × 640. After training for 150 epochs, the experimental results were obtained. The summarized data are presented in Table 4.

According to the table data, the p value of the baseline algorithm for detecting small targets in UAV aerial images is 45.8%, and the Map value is 34.5%. After replacing the backbone network with EfficientNet, the Map improves by 0.3% compared to the baseline algorithm. When the EC2f module is applied, the Map improves by 0.5% relative to the baseline. With the addition of the DTADH module, the overall Map increases by 2.8% compared to the baseline, while maintaining a balanced overall accuracy. Finally, after applying mixed knowledge distillation, the overall p value increases to 48.9%, a 3.1%

improvement, and the Map value reaches 39.1%, marking a 4.6% increase. These results demonstrate that each module contributes significantly to performance improvement.

Table 4. Data sheet of ablation experiment results. A represents the basic YOLOv8n algorithm; B includes the basic YOLOv8n algorithm with the EfficientNet backbone network; C adds the EC2f module to the setup in B; D incorporates the DTADH detection head in addition to the modules in C; and E integrates hybrid knowledge distillation on top of D.

Algorithm	*p* (%)	R (%)	Map 0.5 (%)	Map 0.5: 0.95 (%)	Par (M)	FLOPs (G)
A	45.8	34.0	34.5	20.0	3.17	8.9
B	44.2	34.8	34.8	20.2	7.24	6.2
C	44.8	35.0	35.0	20.3	7.24	6.2
D	45.8	36.8	36.6	21.5	6.28	6.0
E	48.9	38.1	39.1	23.1	6.16	5.4

In summary, compared to the YOLOv8n algorithm, the detection accuracy for small targets in UAV aerial images is 45.8% (*p* value) and 34.5% (Map value). The detection accuracy of the EDANet algorithm achieves a *p* value of 48.9%, representing an increase of 3.1%. The Map value reaches 39.1%, a 4.6% improvement. However, the number of parameters increases by 95%, and the overall computational cost (GFLOPs) is reduced by 39.3%. The optimized algorithm enhances the model's expressive capability through a well-designed architecture. Nevertheless, the algorithm's parameter count has doubled, leading to an increase in overall complexity and storage requirements. Future work could explore the integration of pruning strategies to control the parameter count. Additionally, efficient computing strategies could be applied to further reduce GFLOPs, thus minimizing computational overhead. By optimizing inference speed, while maintaining or improving model accuracy, this optimization is particularly suitable for application scenarios that demand high real-time performance and efficient resource utilization. It not only enhances small target detection accuracy but also improves computational efficiency, significantly increasing the algorithm's feasibility and deployment potential in practical applications.

5.3. Visualization of Experimental Results

To illustrate the role of each module more intuitively, Figure 5 presents the effects as visualized through attention heat maps. The heat map analysis reveals the following: Figure 5b shows an aggregation of small targets but also highlights interference from roofs and trees. This demonstrates a significant challenge in small object detection tasks, where background elements can obscure target detection. Figure 5c illustrates the results after processing the input image through the self-attention and convolution modules of the feature extraction network. The effective backbone features separate complex background details from underlying small target features, allowing better focus on densely clustered small targets while mitigating background interference. Figure 5d displays the outcome after applying the EC2f module for detailed feature extraction. It enhances the perception of small target details, emphasizing the algorithm's focus on the primary target area. Figure 5e shows that EDANet can accurately perceive and locate small targets that were previously affected by background interference. The EC2f module's multi-dimensional feature extraction efficiently utilizes low-level details, reducing the impact of complex backgrounds and occlusions and improving sensitivity and robustness to small targets.

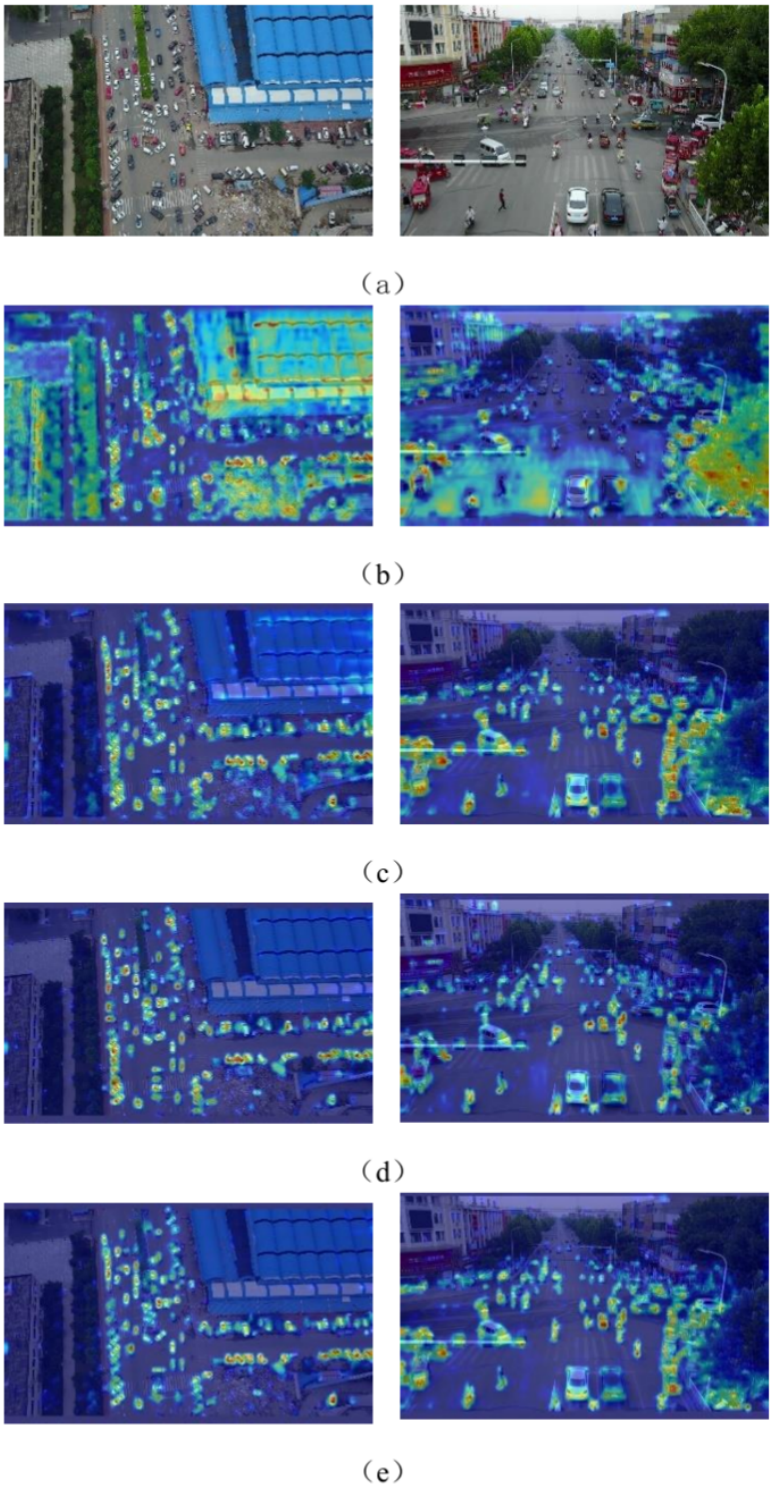

Figure 5. The heat map of the EDANet module is displayed. (**a**) Represents the original picture entered in the algorithm; (**b**) represents the visual display of the heat map of the input picture; (**c**) represents the visualization result of the image heat map after efficient trunk replacement; (**d**) represents the visualization result of the heat map after processing by the EC2f module; and (**e**) represents the visualization result of the heat map after processing by the EC2f module. Visualization results of the heat map processed by the DTADH module.

A comprehensive analysis reveals that the EfficientNet backbone network dynamically adjusts the network structure size of the entire algorithm through compound scaling, enabling it to adapt to the application environment for small targets and efficiently ex-

tract underlying detailed features in complex environments. Simultaneously, the EC2f module performs feature interaction modeling across multiple dimensions, facilitating the multi-dimensional extraction of low-level detail features of small targets in environments with significant background interference and tree cover, thereby mitigating the impact of complex backgrounds and obstructions. This approach enhances the model's sensitivity and robustness to small targets. The DTADH detection module employs information fusion via the feature fusion network, allowing for improved focus on parallel classification and positioning tasks, ultimately leading to more accurate target detection.

To assess the effectiveness of the improved algorithm in real-world scenarios, we used the VisDrone2019 test set, which includes images with small targets, complex backgrounds, and high detection difficulty. The actual detection performance of both YOLOv8 and EDANet was evaluated at different spatial heights. The results are visualized in Figure 6. By comparing the detection results in Figure 6a,b, it is evident that in scenarios where small targets are densely aggregated, the basic algorithm experiences significant clustering and a high rate of missed detections due to overlapping targets. In contrast, EDANet successfully re-detects targets that were previously obscured. Figure 6c,d illustrate UAV images captured from high altitudes, where the wide field of view introduces complex small target imaging and severe background interference, thereby increasing detection difficulty. Under these conditions, EDANet effectively mitigates missed and false detections. Figure 6e,f present UAV images captured from low and medium altitudes, where dense buildings and houses create severe background interference. Despite these challenging imaging conditions, EDANet demonstrates its capability to accurately identify small targets.

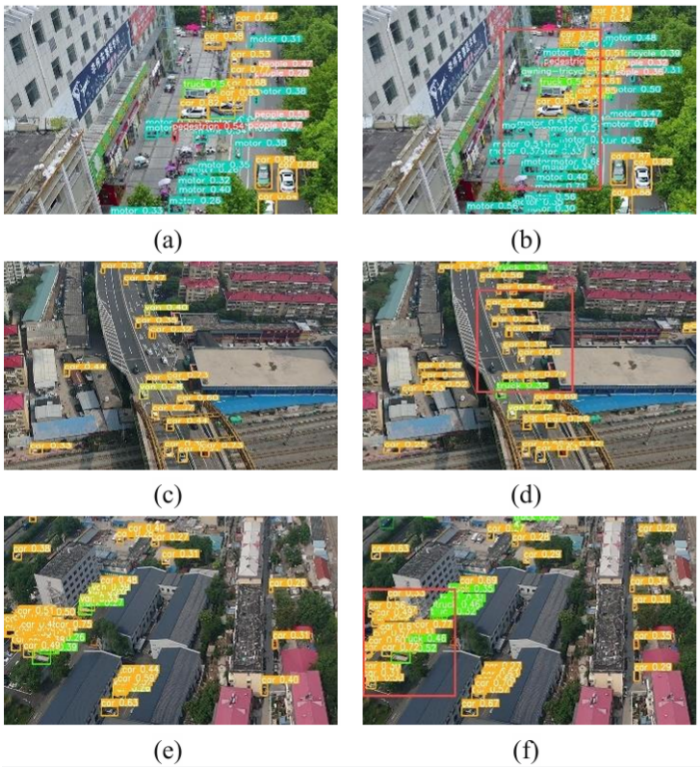

Figure 6. Visual comparison chart of improved algorithms. The left image illustrates the detection performance of YOLOv8, while the right image depicts the detection performance of EDANet. (**a**,**b**) Represent the detection effect of the two algorithms in the dense state of the small target; (**c**,**d**) represent the comparison of the detection effect of the two algorithms in the high-altitude state of the small target; and (**e**,**f**) represent the comparison of the detection effects of the two algorithms when the small target is densely presented and obscured by the architectural background. The red box highlights the area with prominent contrast for small-target inspection.

Given the challenges posed by varying fields of view, target aggregation, the small proportion of target pixels, and susceptibility to background interference, missed detection remains a significant issue. EDANet addresses this by re-detecting and identifying missed targets in key areas, effectively demonstrating the efficacy of the proposed module for small target detection.

6. Conclusions

This paper presents an algorithm aimed at small target detection in UAV aerial images, addressing the challenges of low resolution and difficulty in detection due to background interference. By replacing the existing backbone network with the more efficient EfficientNet, a compound scaling factor is introduced to adjust the depth, width, and resolution of the backbone, thereby enhancing its adaptability to small target detection tasks. This adjustment enables efficient feature extraction and improves feature representation capabilities. Simultaneously, the optimized EC2f module is designed for cross-dimensional interaction modeling, encoding global information through parallel networks and capturing fine-grained features for channel weight optimization. This reduces computational overhead while retaining critical channel information.

The DTADH module performs feature interaction by constructing an efficient multi-level shared feature network. The output features are dynamically assigned, and task-specific interaction features are learned across multiple convolutional layers through a task alignment predictor, enabling joint feature maps that are better aligned for both classification and positioning tasks. The depth and number of feature layers can be flexibly adjusted to better match the requirements of classification and positioning tasks. Additionally, hybrid knowledge distillation is applied to optimize the overall detection performance. Soft labels and feature maps are combined to construct loss distillation functions for classification and positioning, facilitating logical distillation. The feature loss function, derived from the differences between various feature maps, enhances the student model's ability to recognize small targets, improving detection accuracy and boosting the generalization capability of the algorithm in small target detection scenarios.

The comprehensively improved algorithm significantly enhances the accuracy of small target detection while effectively reducing computational complexity. This design holds substantial practical value for UAV-based edge devices in street traffic information acquisition and traffic monitoring tasks in complex environments. Although the algorithm's average accuracy improvement is relatively modest and the number of parameters has increased, the detection performance in complex scenes has been notably improved. Future research should focus on further lightweighting the model to better suit resource-limited edge devices. Additionally, efforts will be directed towards reducing the number of parameters and developing a more efficient detection module, while also improving the algorithm's robustness by simulating more real-world scenarios. Furthermore, quantitative training or pruning techniques will be applied to EDANet to address the needs of practical applications and achieve more efficient detection results.

Author Contributions: F.Z. led the project design, proposed the core ideas, and developed the model improvements in detail. G.Z. was responsible for creating the experimental program, which included selecting the dataset, conducting model training and testing, and drafting the manuscript. Z.W. contributed to this study's conceptual framework, the design, and performed the statistical analysis of the findings. S.Y. and Z.L. oversaw the project, ensuring the rigor of the experimental process and the reliability of the results. All authors reviewed and discussed the results, provided feedback on the manuscript, and gave final approval. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Key-Area Research and Development Program of Guangdong Province under Grant 2021B0101420001, Jiangxi Provincial Natural Science Foundation No. 20232ABC03A07, and China National Railway Group Co., Ltd. Science and Technology Research and Development Program Project L2022X002.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The author Shengli Yang was employed by the company CHN Energy Investment Group Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GFLOPs	Giga Floating-Point Operations per second
Map	Mean Average Precision
TP	True Positive
FP	False Positive
FN	False Negative

References

1. Wang, L.; Liu, J.; Wang, W. Small target detection method in UAV images based on dilated convolution fusion Transformer. *J. Comput. Appl.* **2024**, *18*, 1–14.
2. Liu, P.; Qian, W.; Wang, Y. YWnet: A convolutional block attention-based fusion deep learning method for complex underwater small target detection. *Ecol. Inform.* **2024**, *79*, 102401. [CrossRef]
3. Zhang, F.; Lin, S.; Xiao, X.; Wang, Y.; Zhao, Y. Global attention network with multiscale feature fusion for infrared small target detection. *Opt. Laser Technol.* **2024**, *168*, 110012. [CrossRef]
4. Sun, L.; Cai, Z.; Liang, K.; Wang, Y.; Zeng, W.; Yan, X. An intelligent system for high-density small target pest identification and infestation level determination based on an improved YOLOv5 model. *Expert Syst. Appl.* **2024**, *239*, 122190. [CrossRef]
5. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
6. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef]
7. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
8. Zhu, S.; Gao, H. MC-ShuffleNetV2: A lightweight model for maize disease recognition. *Egypt. Inform. J.* **2024**, *27*, 100503. [CrossRef]
9. Zhao, L.; Wang, L. A new lightweight network based on MobileNetV3. *KSII Trans. Internet Inf. Syst. (TIIS)* **2022**, *16*, 1–15.
10. Li, Z.; Wang, Z.; He, Y. Aerial photography dense small target detection algorithm based on adaptive collaborative attention mechanism. *Acta Aeronaut. Astronaut. Sin.* **2023**, *168*, 110012.
11. Hui, Y.; Wang, J.; Li, B. STF-YOLO: A small target detection algorithm for UAV remote sensing images based on improved SwinTransformer and class weighted classification decoupling head. *Measurement* **2024**, *224*, 113936. [CrossRef]
12. Yin, N.; Liu, C.; Tian, R.; Qian, X. SDPDet: Learning Scale-Separated Dynamic Proposals for End-to-End Drone-View Detection. *IEEE Trans. Multimed.* **2024**, *26*, 7812–7822. [CrossRef]
13. Xue, C.; Xia, Y.; Wu, M.; Chen, Z.; Cheng, F.; Yun, L. EL-YOLO: An efficient and lightweight low-altitude aerial objects detector for onboard applications. *Expert Syst. Appl.* **2024**, *256*, 124848. [CrossRef]
14. Yang, B.; Zhang, X.; Zhang, J.; Luo, J.; Zhou, M.; Pi, Y. EFLNet: Enhancing Feature Learning Network for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5906511. [CrossRef]
15. Tang, X.; Deng, H.; Liu, G.; Li, G.; Li, Q.; Zhao, J.; Zhou, Y. FR-YOLOv7: Feature Enhanced YOLOv7 for Rotated Small Object Detection in Aerial Images. *Meas. Sci. Technol.* **2024**, *35*, 116004. [CrossRef]
16. Qin, L.; Pang, W.; Zhao, D. A feature pyramid network with adaptive fusion strategy and enhanced semantic information. *Multimed. Syst.* **2024**, *30*, 171. [CrossRef]

17. Lau, K.W.; Po, L.M.; Rehman, Y.A.U. Large Separable Kernel Attention: Rethinking the Large Kernel Attention Design in CNN. *Expert Syst. Appl.* **2023**, *236*, 121352. [CrossRef]
18. Min, X.; Zhou, W.; Hu, R.; Wu, Y.; Pang, Y.; Yi, J. Lwuavdet: A lightweight uav object detection network on edge devices. *IEEE Internet Things J.* **2024**, *11*, 24013–24023. [CrossRef]
19. Li, X.; Wang, F.; Wang, W.; Han, Y.; Zhang, J. DM-YOLOX aerial object detection method with intensive attention mechanism. *J. Supercomput.* **2024**, *80*, 12790–12812. [CrossRef]
20. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
21. Jiang, L.; Yuan, B.; Du, J.; Chen, B.; Xie, H.; Tian, J.; Yuan, Z. MFFSODNet: Multi-Scale Feature Fusion Small Object Detection Network for UAV Aerial Images. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 5015214. [CrossRef]
22. Buciluǎ, C.; Caruana, R.; Niculescu-Mizil, A. Model compression. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 20–23 August 2006; pp. 535–541.
23. Hinton, G.E.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *arXiv* **2015**, arXiv:1503.02531.
24. Romero, A.; Ballas, N.; Kahou, S.E.; Chassang, A.; Gatta, C.; Bengio, Y. FitNets: Hints for Thin Deep Nets. *arXiv* **2014**, arXiv:1412.6550.
25. Fan, Q.; Huang, H.; Chen, M.; Liu, H.; He, R. Rmt: Retentive networks meet vision transformers. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 5641–5651.
26. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv* **2019**, arXiv:1905.11946.
27. Feng, C.; Zhong, Y.; Gao, Y.; Scott, M.R.; Huang, W. Tood: Task-aligned one-stage object detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 3490–3499.
28. Tian, Z.; Chu, X.; Wang, X.; Wei, X.; Shen, C. Fully convolutional one-stage 3d object detection on lidar range images. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 34899–34911.
29. Du, D.; Zhu, P.; Wen, L.; Bian, X.; Lin, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y.; et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019; pp. 213–226.