remote sensing

MDPI

Article

YOLO-ViT-Based Method for Unmanned Aerial Vehicle Infrared Vehicle Target Detection

Xiaofeng Zhao, Yuting Xia *, Wenwen Zhang, Chao Zheng and Zhili Zhang

Xi'an Research Institute of High-Tech, Xi'an 710025, China; xife_zhao@outlook.com (X.Z.)
* Correspondence: xiayt1@outlook.com (Y.X.)

Abstract: The detection of infrared vehicle targets by UAVs poses significant challenges in the presence of complex ground backgrounds, high target density, and a large proportion of small targets, which result in high false alarm rates. To alleviate these deficiencies, a novel YOLOv7-based, multi-scale target detection method for infrared vehicle targets is proposed, which is termed YOLO-ViT. Firstly, within the YOLOV7-based framework, the lightweight MobileViT network is incorporated as the feature extraction backbone network to fully extract the local and global features of the object and reduce the complexity of the model. Secondly, an innovative C3-PANet neural network structure is delicately designed, which adopts the CARAFE upsampling method to utilize the semantic information in the feature map and improve the model's recognition accuracy of the target region. In conjunction with the C3 structure, the receptive field will be increased to enhance the network's accuracy in recognizing small targets and model generalization ability. Finally, the K-means++ clustering method is utilized to optimize the anchor box size, leading to the design of anchor boxes better suited for detecting small infrared targets from UAVs, thereby improving detection efficiency. The present article showcases experimental findings attained through the use of the HIT-UAV public dataset. The results demonstrate that the enhanced YOLO-ViT approach, in comparison to the original method, achieves a reduction in the number of parameters by 49.9% and floating-point operations by 67.9%. Furthermore, the mean average precision (mAP) exhibits an improvement of 0.9% over the existing algorithm, reaching a value of 94.5%, which validates the effectiveness of the method for UAV infrared vehicle target detection.

Keywords: unmanned aerial vehicle target detection; vehicle detection; infrared small target; deep learning; Yolov7

Citation: Zhao, X.; Xia, Y.; Zhang, W.; Zheng, C.; Zhang, Z. YOLO-ViT-Based Method for Unmanned Aerial Vehicle Infrared Vehicle Target Detection. *Remote Sens.* **2023**, *15*, 3778. https://doi.org/10.3390/rs15153778

Academic Editors: Francisco Agüera-Vega, Fernando Carvajal-Ramírez and Patricio Martínez-Carricondo

Received: 14 June 2023
Revised: 13 July 2023
Accepted: 27 July 2023
Published: 29 July 2023

1. Introduction

Over the past few years, small unmanned aerial vehicles (UAVs) have become increasingly prevalent in both civilian and military contexts as an innovative image acquisition platform. This is due to their low cost, high degree of flight flexibility, small size, excellent concealment, and high level of efficiency [1]. One particular area of focus has been the detection of vehicle targets using UAVs, which has garnered significant attention and has been applied in a variety of domains, including traffic vehicle monitoring [2–4], accident search and rescue [5,6], road planning [7], and military intelligence collection [8]. Currently, with the continuous development of deep learning, vehicle detection in UAV-based scenarios as a hot topic in deep learning research has also achieved many good results [9–11]. However, the interference of weather and lighting with visible images, especially in nighttime conditions, poses a significant challenge to vehicle target detection tasks when using only visible light detection [12]. Infrared imaging technology has strong anti-interference ability, long detection distance, and all-weather detection advantages and has been applied to infrared vehicle detection technology based on drone images [13,14]. Therefore, it is of great importance and application to study how to accurately handle various complex scenarios of UAVs vehicle detection technology. Detecting vehicle targets through infrared

technology using UAVs poses implementation challenges. One of the primary obstacles is the variable height of the UAVs and their large camera field of view, which can result in significant variations in the scale of the infrared targets obtained, even for the same target. Moreover, the targets are typically presented as dense small targets, and the lack of color, texture, and other features in infrared images renders small targets vulnerable to interference, thereby increasing the false alarm rate during detection. Consequently, improving the accuracy of infrared vehicle target detection using UAVs has emerged as a prominent research area.

In Microsoft Common Objects in Context (MS COCO) [15], small targets are defined as targets with a resolution of less than 32 × 32 pixels, medium targets are larger than 32 × 32 pixels and smaller than 96 × 96 pixels, and targets larger than 96 × 96 pixels are large targets. In complex backgrounds with significant noise, the detection of small targets in infrared images is severely affected. Compared to traditional algorithms, deep-learning-based object detection algorithms avoid the need for extensive manual design and demonstrate superior performance in detection tasks. At present, deep learning algorithms are classified into two categories. The first category includes two-stage algorithms that are based on candidate regions, such as R-CNN [16] and its derived algorithms [17,18]. Although these algorithms can achieve higher accuracy, they are associated with speed and efficiency issues. Liu Shunmin et al. [19] confirmed the effectiveness of faster R-CNN in remote-sensing target detection on the NWPU VHR remote-sensing dataset. The second category comprises one-stage algorithms such as SSD [20] and YOLO [21–23], which have made specialized improvements to address speed and efficiency challenges. Chen et al. [24] conducted a study on the detection of military and civilian vehicles on the edge devices of drones using quantized SSD and YOLOv3 structures. However, due to the prevalence of small targets and unclear target features in UAVs' aerial image datasets, it is challenging to achieve optimal results when applying general algorithms to infrared vehicle target detection in UAVs.

In response to the problem of low detection accuracy caused by small targets and multiple scales in infrared vehicle target detection of UAVs, this paper proposes a YOLO-ViT-based method for infrared vehicle target detection of UAVs. Through the K-Means++ algorithm, the anchor box size is adjusted, and the lightweight MobileViT network is used to replace the original backbone network, reducing the number of model parameters while fully extracting the local and global features of the target. The C3-PANet structure is proposed, which combines the feature recombination upsampling method CARAFE and the C3 structure to reduce the number of model parameters while increasing the feature receptive field, fully utilizing feature information, and improving vehicle detection performance. The main work of this paper is as follows:

1. In this paper, we use the lightweight network MobileViT as the backbone network, which combines the spatial inductive bias of CNNs and the global modeling capability of transformers, reducing the model parameters and complexity.
2. The multi-scale C3-PANet structure is proposed, which uses the feature recombination and upsampling CARAFE method to predict the upsampling weights based on the feature map, and then recombines the features based on the predicted upsampling weights to obtain a larger receptive field and enhance the perception ability of small targets. The neck structure is improved by using the C3 structure and by stacking C3 modules to extract more effective features while reducing the number of parameters and improving the detection accuracy of small targets.
3. A K-means++ clustering algorithm is introduced to cluster the dataset samples and redesign the anchor box size to improve detection efficiency.

2. Related Work

UAVs exhibit multi-scale, multi-angle, and multi-vehicle characteristics, which pose greater challenges for vehicle detection than ordinary road vehicles. To overcome these challenges, researchers have proposed numerous enhancement methods that rely on UAV

scenes. These methods aim to improve the performance of network target detection and reduce false alarm rates for small targets. To this end, various techniques have been employed, including the redesign of anchor boxes and feature extractors, the utilization of data augmentation techniques, and a combination of different deep learning methods. As a consequence, more appropriate vehicle detection tasks for UAV scenes have been proposed, resulting in a significant improvement in the detection accuracy of vehicle targets.

The prior box is a predefined size box that predicts the boundary box of the target. Through anchor boxes of different sizes, it can adapt to targets of different scales. Many studies have shown that the size of anchor boxes has a crucial impact on vehicle detection [17–19,25]. Chen et al. [24] proposed a method based on SSD, which adjusts the size of prior boxes to improve the detection accuracy of small targets. Mulan Qiu et al. [26] proposed a drone road detection method ASFF-YOLOv5 based on multi-scale feature fusion. By analyzing with the K-means++ clustering algorithm, the anchor box size is reconstructed, and the ASFF detection head and SPPF spatial pyramid pooling structure are used to improve feature scale invariance and enhance the target detection effect. To address the impact of multi-scale on detection, Liu et al. [27] proposed a method called CAFFNet for detecting traffic signs, which introduces a channel attention and feature fusion multi-objective detection method, utilizing multi-scale contextual information to reduce feature differences and improve detection accuracy. The model is trained using class pairs of different feature scales and through a multi-scale fusion algorithm. In the loss function, Gaussian distribution features are applied to enhance detection accuracy. Liu [28] presented a dense-connection-based multi-scale fusion method called DMFFPN, which fully utilizes the features of each convolutional layer through dense connections, adopts a cascade architecture to enhance local generalization ability, and achieves good results on the VisDrone dataset [29]. Sun et al. [30] developed the I-YOLO model, which incorporates EfficientNet to improve the original network structure and enhance feature extraction. Additionally, the U-Net residual network is utilized to reduce infrared noise, and the K-means algorithm is employed to anchor box sizes, thereby improving the model's ability to detect road vehicles.

In order to improve the feature information extraction of small targets in models, Tang et al. [31] introduced the hyper-RPN technique. This method combines feature maps from various layers to effectively detect small vehicle targets by merging shallow and deep feature information. Another approach to address this issue was proposed by Zhu [32], who developed TPH-YOLOv5. This method incorporates a small target detection head and integrates a transformer to enhance the structure of the remaining detection heads, resulting in improved accuracy of target detection. In their research, Zuo et al. [33] presented a pyramid structure called AFFPN that utilizes attention feature fusion. They developed attention fusion modules that enhance the position and semantic information of shallow and deep layers, resulting in improved target feature extraction and excellent performance on public datasets. Yao et al. [34] introduced the FCOS model, which leverages improved spatial feature fusion and traditional filtering methods to suppress background noise and improve the detection of small targets. Additionally, Zhang et al. [35] proposed a novel infrared target detection method known as CHFNet. This method employs a local fusion HLF module as a cross-layer feature-fusion module, which effectively reduces the loss of feature information and enhances the detection performance of weak infrared targets. Li et al. [36] proposed the YOLO-FIRI model to improve the structure of the feature extraction network and to use the multi-scale structure to improve the detection accuracy of infrared small target detection. Dai et al. [37] proposed an infrared small target detection method and designed an asymmetric contextual modulation module (ACMM) to better extract target features. Zhang et al. [38] proposed ISNet and designed the Taylor finite difference (TFD)-inspired edge block and two-orientation attention aggregation (TOAA) block, which efficiently extracts weak infrared targets from blurred backgrounds with edge features from the blurred background.

According to the introduction of the transformer [39] in NLP, researchers have been exploring ways to apply it to computer vision. According to studies [40–42], the transformer model has shown promising results in computer vision, opening up new possibilities for the field. One such application is object detection, which Carion et al. [43] simplified through their proposal of the DETR model. In addition, Dosovitskiy et al. [40] achieved good results in image classification tasks with the ViT model, particularly when trained with large amounts of data. In the domain of target detection, the transformer model has been widely employed. In this regard, Liu et al. [44] incorporated an attention mechanism into the network, which helped to enhance the receptive field and capture relevant information regarding small infrared targets. On the other hand, Chen et al. [45] developed the IRSFormer model that utilizes the transformer structure to extract features, encodes the infrared images using HOSPT, and fuses feature information through a top-down multi-scale fusion module TFAM structure. As a result, the detection efficiency of small infrared targets has been significantly improved. Touvron et al. [46] proposed the DeiT model, which compresses the ViT model by knowledge distillation, using soft distillation and hard-label distillation methods by distilling the soft labels output from the teacher model, and the actual labels predicted by the teacher network, respectively. After the distillation of knowledge from the model, DeiT has fewer parameters and faster inference and achieves good results. Rao et al. [47] proposed the DynamicViT model, which achieves speedup by sparsifying the inputs of each layer in the transformer model and removing a portion of redundant parameters, achieving good results.

3. Proposed Methodology, Tools, and Techniques

This section is divided into three parts, which introduce the general structure and principles of the YOLO-ViT model, the high-altitude infrared dataset HIT-UAV [48,49], and the evaluation metrics used to validate the methodology of this paper.

3.1. YOLO-ViT

This article proposes a YOLO-ViT-based UAV infrared vehicle target detection method, as shown in Figure 1. The backbone network of the YOLO-ViT model uses the MobileViT network structure, combined with the feature extraction capabilities of CNNs and transformers, to generate multi-scale feature maps with richer details. The neck network adopts the C3-PANet multi-scale feature-fusion structure, mainly composed of SPPCSPC module, CARAFE upsampling module, C3 structure, and MP downsampling, which aggregates upper- and lower-layer information through a bottom-up path, fuses feature maps that retain more details and improves the accuracy of vehicle detection. Finally, the obtained multi-scale feature maps are used to obtain prediction results through the detection head.

SPPCSPC represents a pyramidal pooling structure. The lack of feature information in IR images, such as texture and color, when compared to visible images, causes MP. To address this issue and to ensure that the features of IR images are fully utilized while minimizing missed and false detections in detection results, the upsampling CARAFE module with feature reorganization may be employed. This approach allows for better utilization of information surrounding the features, increases the number of parameters and calculations, and introduces only a limited number of the perceptual field of the feature map, thus improving the feature extraction of small IR targets. Additionally, the model's power is enhanced by improving the bottleneck structure and introducing a C3 structure to improve the model's characterization. This approach lightens the neck structure while enhancing the extraction and fusion of features.

3.1.1. Improved Backbone Network Based on MobileViT

In recent years, transformer models have gained popularity in various natural language processing tasks owing to their ability to capture global information interactions. However, these models tend to generate an excessive number of parameters during training, rendering them unsuitable for deployment on mobile hardware. As a solution to this

problem, Apple proposed the MobileViT [50] network in 2021. This lightweight, general-purpose, mobile-friendly transformer model combines the spatialized paranoid induction capabilities of CNNs with the global processing ability of transformers. By doing so, it can effectively learn both local and global features of an image while utilizing fewer parameters and simpler training methods. This, in turn, improves the accuracy of target detection. In light of these advantages, we have chosen to use the MobileViT network as the backbone network for YOLO-ViT.

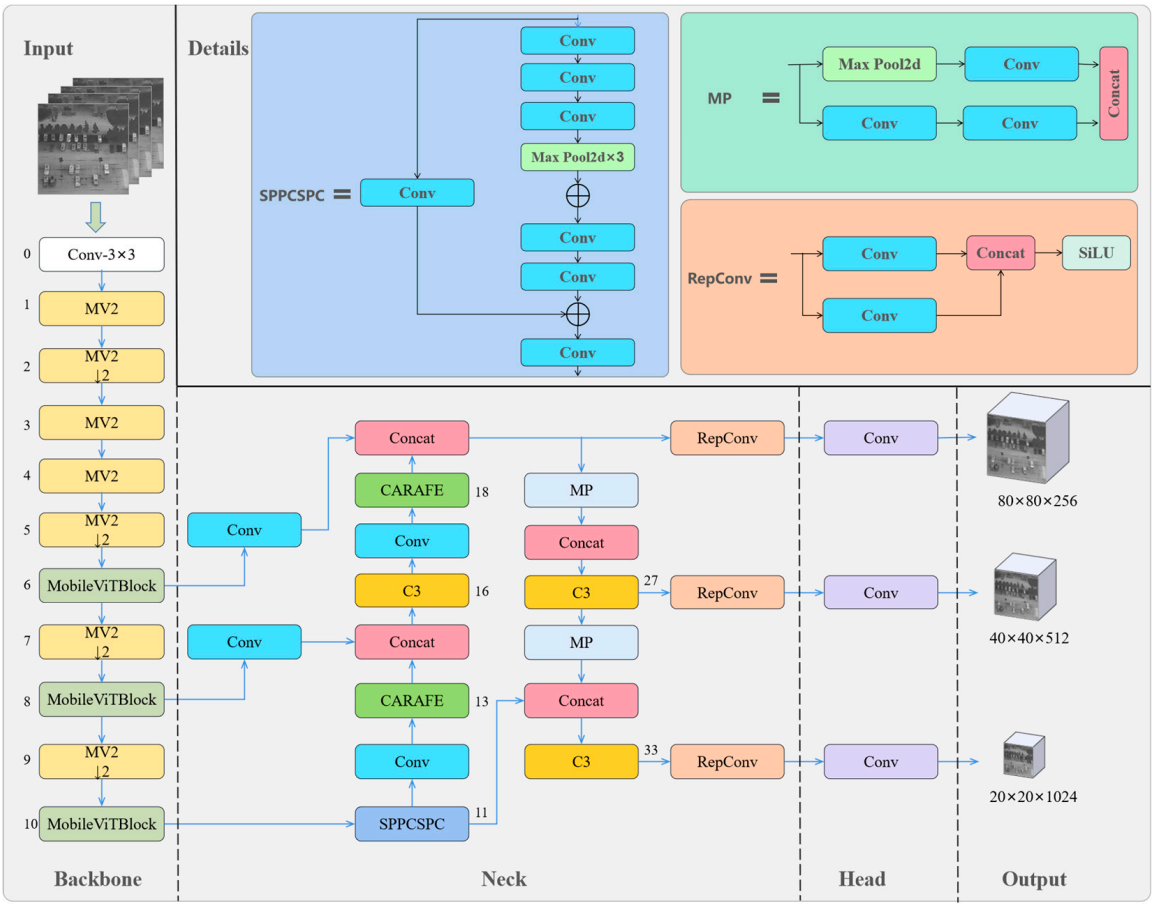

Figure 1. The YOLO-ViT network structure diagram. SPPCSPC, MP, and RepConv are three models in the neck of the YOLOv7, and the details of these models are shown in the upper half of the figure. SPPCSPC is a pyramid pooling structure, MP is a downsampling structure, and RepConv is a multiparameter structure.

The MobileViT architecture is primarily composed of two distinct blocks, namely MV2Block and MobileViTBlock. The former is derived from the inverted residual block (IRB) structure that was originally introduced in MobileNetv2 [51]. This block structure, as illustrated in Figure 2, involves increasing the dimensionality of the feature map through 1×1 convolution, followed by processing using 3×3 depth separable convolution. The MVBlock module serves to minimize the loss of feature information and enhance the extraction of target features by modifying the dimensionality of the features in a manner that first increases and then decreases it.

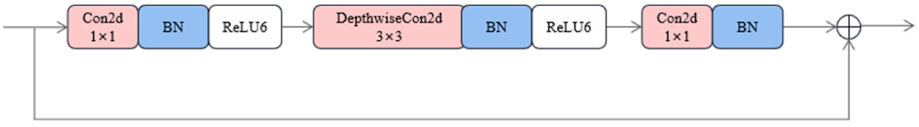

Figure 2. The MV2Block model structure diagram.

The MobileViTBlock unit comprises three key components: a local information encoding module, a global information encoding module, and a feature-fusion module, as illustrated in Figure 3. Initially, the feature map is reconstructed locally via 3 × 3 convolution, and subsequently, the number of channels is adjusted through the convolution of the feature map. Following this, the global information encoding module is engaged, wherein the global information is modeled for the features. In the third step, the number of channels is adjusted back to the original input size by using a convolution layer and then stitched with the feature map of the original input along the channel direction. Eventually, the output fused feature map is obtained through the 3 × 3 convolution of the features. This technique is capable of reducing computational effort while completely extracting the local and global feature information.

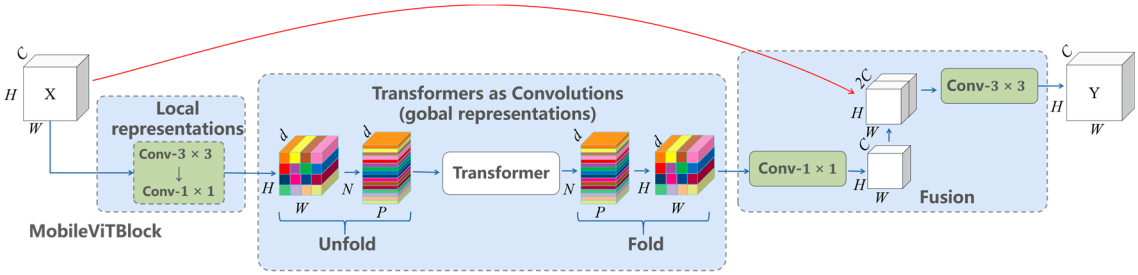

Figure 3. The MobileViTBlock model structure diagram.

This study utilizes the MobileViT network as the foundation of the YOLO-ViT model, wherein the global pooling and classification layers are omitted. The quantity of network channels is documented in Table 1. The terminology used in Table 1 is defined as follows: layer refers to the number of layers in the network; output size pertains to the resultant image size; stride denotes the displacement of the convolutional kernel; number denotes the number of modules; output channels refers to the number of input channels; ↓2 represents downsampling. and L represents the number of transformer repetitions present in the MobileViT architecture.

Table 1. The parameters of the MobileViT structure.

Layer	Output Size	Stride	Number	Output Channels
Image	640 × 640	1		
Conv-3 × 3, ↓2	320 × 320	2	1	16
MV2			1	32
MV2, ↓2	160 × 160	4	1	48
MV2			2	48
MV2, ↓2	80 × 80	8	1	64
MobileViTBlock (L = 2)			1	64
MV2, ↓2	40 × 40	16	1	80
MobileViTBlock (L = 4)			1	80
MV2, ↓2	20 × 20	32	1	96
MobileViTBlock (L = 3)			1	96

3.1.2. Content-Aware Multi-Scale-Structure-Based C3-PANet

The detection of infrared vehicle targets by UAVs in complex backgrounds is a challenging task due to the multi-scale characteristics and high prevalence of small targets. The good maneuverability of UAVs and the large field of view of aerial images offer potential solutions to this problem. However, the lightweight MobileViT backbone network, while reducing the model's weight, can compromise the accuracy of detection. To overcome these difficulties, we propose a multi-scale fusion network structure, C3-PANet, which op-

timizes the information propagation path by shortening it through a bottom-up approach, merges multi-scale feature information, improves network performance using the CARAFE upsampling operator and C3 structure, increases the perceptual field of the feature map, and reduces the loss of feature information for small targets. By improving the perceptual field of the feature map and reducing the loss of feature information for small targets, our proposed network structure enhances the accuracy of detecting IR vehicle targets using UAV-based systems.

The process of upsampling in CARAFE involves the use of an upsampling kernel, which is applied to each position of the feature map, and then the dot product of pixels in the corresponding neighborhood is taken from the input feature map. This technique is also referred to as feature recombination. Traditional methods for upsampling, such as nearest neighbor interpolation and bilinear interpolation, simply insert new elements between pixels of the original image to increase the number of pixels without utilizing the semantic information of the feature map. As a result, the obtained feature map has a small perceptual field, and the feature information of the UAV infrared vehicle target is easily lost. Wang et al. [52] proposed a novel upsampling method named CARAFE, which predicts the size of the upsampling kernel from the input feature content while introducing only a small number of additional parameters and computational effort. By using the predicted results to guide the upsampling process, a larger field of perception can be achieved during the upsampling operation, and the semantic information of the features can be fully utilized to extract the target features.

The CARAFE upsampling operator is divided into two parts, the convolution kernel prediction module and the feature recombination module, and the structure is shown in Figure 4. In the convolution kernel prediction module, the size of the upsampled convolution kernel is predicted by the convolution kernel prediction module for the target location in the input feature map. The upsampling multiplicity is assumed to be σ, and the input feature map χ is $H \times W \times C$. Firstly, the channel C is compressed to C_{m} by 1×1 convolution to obtain a feature map $\chi\prime$ of size $H \times W \times C$, which reduces the computational effort of subsequent operations. Secondly, the input feature map $\chi\prime$ is encoded into the encoder and the upsampling kernel $w_{l\prime}$ is predicted by a convolutional layer of $k_{encoder} \times k_{encoder}$, and the number of output channels is $\sigma^2 k_{up}^2$, which is expanded to obtain an upsampling kernel of size $\sigma H \times \sigma W \times k_{up}^2$. Finally, the softmax function is used to normalize the predicted upsampling kernels so that the convolutional kernel weights sum to 1.

In the feature reorganization module, the obtained upsampled kernel is used for feature recombination to extract the target features. For each target position of the output feature map, it is mapped back to the input feature map by taking a region of size $k_{up} \times k_{up}$ centered on the target and making a dot product with the upsampling kernel obtained from the prediction of the position of that point, resulting in a feature map of size $\sigma H \times \sigma W \times C$. The calculation method is as follows:

$$\chi\prime_{l\prime} = \sum_{n=-r}^{r} \sum_{m=-r}^{r} w_{l\prime(n,m)} \cdot \chi_{(i+n,j+m)} \tag{1}$$

where (i, j) denotes the location of the point, $r = k_{up}/2$. Using a content-aware upsampling-based approach, it is possible to obtain feature maps containing semantic information, focusing more on features from local regions and improving the detection accuracy of small targets.

The C3 module is a suggested framework in the YOLOv5 network that aims to augment the depth of the network and the perceptual field of the feature map while also enhancing the capacity to extract features. This module comprises three Conv modules and one BottleNeck module, as illustrated in Figure 5. The BottleNeck module leverages a residual connection to decrease the dimensionality of the feature map by means of the convolution operation, thus attaining global spatial information. Ultimately, a residual

structure is used to sum the input and output, which mitigates the issue of gradient disappearance.

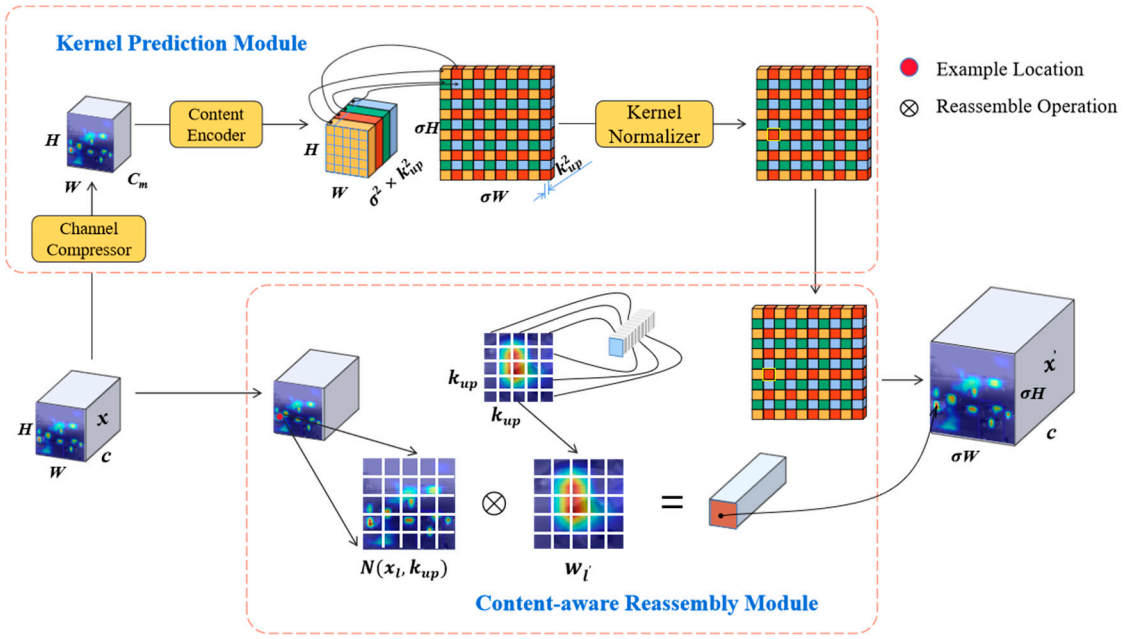

Figure 4. CARAFE structure diagram.

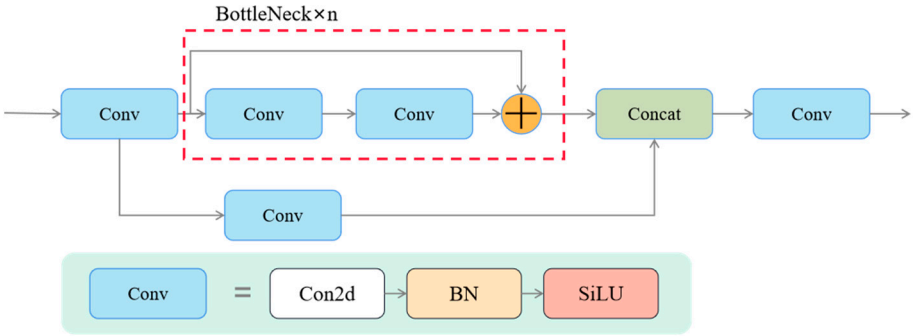

Figure 5. C3 module structure diagram.

3.1.3. K-Means++ Clustering Algorithm

The K-means clustering algorithm is a type of centroid-based clustering that groups samples iteratively into classes. This process is carried out in such a way that the sum of distances between each sample and the midpoint or mean of its respective class is minimized. Distance, in this context, is used to measure the similarity or difference between samples, with smaller distances indicating more similarity between samples and larger distances indicating more difference. However, the algorithm's convergence is heavily reliant on the initialization of cluster centers, which is random. To address this issue, the K-means++ clustering algorithm [53] has been developed. This algorithm improves the initialization process by randomly selecting a point as the initial center and then selecting the next point based on its furthest distance from the previous point until an initialized centroid is completed. Following this, cluster analysis is performed.

The detection of targets produces various bounding boxes with different scaling ratios and aspect ratios, each with a pixel as the center. These boxes, referred to as anchor boxes, can affect the accuracy of the model when they are too large or too small. Small targets are typical of UAV aerial images and often exhibit characteristics such as occlusion and multi-scale. Oversized anchor boxes can result in the loss of targets. To improve the ability to learn the position and size of infrared targets, the K-means++ algorithm is utilized to

cluster a dataset of UAV infrared vehicle targets. This generates a dataset that is more appropriate for multi-scale UAV infrared vehicle targets. The resulting dataset is expected to be of greater utility for infrared target detection.

First, a sample point is randomly selected from the data set χ as the first initial cluster center c_i, then the shortest distance $D(x)$ between each sample and the sample point is calculated, then the probability $P(x)$ of each sample point being the next cluster center is calculated, and the probability is calculated as follows:

$$P(x) = \frac{D(x)^2}{\sum_{x \in \chi} D(x)^2} \tag{2}$$

The maximum probability is selected, and the corresponding sample points are used as the next class of centers, iterating until k clustering centers are generated. In this paper, the HIT-UAV dataset was selected for cluster analysis to obtain nine anchor frames {8, 16}, {9, 23}, {4, 22}, {16, 35}, {23, 27}, {31, 46}, {39, 28}, {51, 68}, and {85, 106}.

3.2. Datasets

This paper utilizes the YOLO-ViT method to assess its efficacy in detecting small infrared targets of vehicles in UAV applications. The HIT-UAV dataset is chosen for comparative experiments due to its public availability and comprehensive coverage. This dataset represents a notable contribution to the field as it is the first high-altitude UAV infrared dataset that includes images captured at different altitudes, viewing angles, and object types in Figure 6. The dataset consists of 2898 thermal infrared images with a resolution of 640 × 512 and contains 24,899 labels in five categories, namely Person, Car, Bicycle, OtherVehicle, and DontCare. To simplify the dataset, the DontCare category was removed, and the Car and OtherVehicle categories were merged into a single Vehicle category. This resulted in three categories, namely Person, Vehicle, and Bicycle, with a total of 2866 images.

Figure 6. The HIT-UAV dataset.

In this paper, the dataset was divided into three parts using a ratio of 7:2:1. This resulted in a train set of 2008 images, a test set of 571 images, and a validation set of 287 images. According to the information provided in Table 2, the final dataset consists of 17,118 labels for small targets that are below 32 × 32 pixels in size, 7249 labels for medium targets below 96 × 96 pixels, and 384 labels for large targets. It is worth noting that the smallest target in the HIT-UAV dataset only occupies 0.01% of the image pixels. This

dataset fulfills the criteria for detecting high-altitude small infrared targets using unmanned aerial vehicles.

Table 2. Number of tags for small, medium, and large targets in HIT-UAV dataset.

	Small (0, 32 × 32)	**Medium (32 × 32, 96 × 96)**	**Large (96 × 96, 640 × 512)**
HIT-UAV	17,118	7249	268
Train set	12,045	5205	268
Test set	3331	1379	70
Validation set	1742	665	46

3.3. Assessment Indicators

This article utilizes a selection of evaluation metrics, namely precision (P), recall (R), F1 (F1-score), AP (average precision), mAP (mean average precision), parameters, GFLOPs, and FPS. The F1-score is calculated as a weighted average of confidence and recall, while both AP and mAP serve as ultimate evaluation metrics that gauge the detection accuracy of the model. The model size and algorithmic complexity are measured by the parameters and GFLOPs, respectively. The evaluation parameter equations are provided below.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{4}$$

$$\text{F1-score} = \frac{2 * \text{P} * \text{R}}{\text{P} + \text{R}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}, \tag{5}$$

$$\text{AP} = \int_0^1 \text{P}(\text{r})\text{dr}, \tag{6}$$

$$\text{mAP} = \frac{\sum_{\text{i}=1}^{\text{N}} \text{AP}_\text{i}}{\text{N}}. \tag{7}$$

In the context of classification, the terms true positive (TP), false positive (FP), and false negative (FN) are commonly used to refer to samples that are correctly identified, incorrectly identified, and missed, respectively. N is a variable that represents the number of categories being classified.

4. Experimental Results

4.1. Experimental Platform and Parameter Settings

In order to evaluate the effectiveness of the YOLO-ViT model, a comparison test and an ablation experiment were designed. The hardware platform settings applied during these experiments have been presented in Table 3.

Table 3. Experimental platform configuration.

Names	**Related Configurations**
Graphics processing unit	NVIDIA Quadro GV100
Central processing unit	Inter Xeon Platinum 8151+++
GPU memory size	32 G
Operating system	Win 10
Computing platform	CUDA10.2
Deep learning framework	Pytorch

4.2. Comparison Algorithms

In order to evaluate the usability and efficacy of YOLO-ViT for detecting infrared vehicles in UAV scenarios, a comparative study is conducted with various advanced algorithms under identical conditions. All experimental models were trained without pre-training weights, and training was recommenced. The input image size of the model was adjusted to 640 × 640, with the batch size set at 16 and the number of iterations at 300 epochs. The training results indicate a 0.9% increase in mAP for YOLO-ViT in contrast to the original YOLOv7s, along with a 0.5% increase in AP for detecting the vehicle target, and a 52.6% decrease in parameters and 69.3% decrease in computational effort. With a significantly lower model complexity, YOLO-ViT achieves higher accuracy in detecting UAV infrared vehicle targets. These findings validate the effectiveness of the proposed method for infrared vehicle detection in UAV-based scenarios. The outcomes of this study are presented in Table 4.

Table 4. Performance comparison table between YOLOViT and YOLOv7.

	Parameters	GFLOPs	Precision	Recall(%)	F1(%)	$AP_{Vehicle}$ (%)	mAP(%)
YOLOv7s	36.5 M	103.2	90.2	88.4	89.3	97.6	93.6
YOLO-ViT	**17.3 M**	**33.1**	**90**	**91.3**	**90.6**	**98.1**	**94.5**

To further validate the reliability of YOLO-ViT, we chose other models from the YOLO series as the baseline and conducted experiments on the HIT-UAV UAV infrared target dataset. This dataset acquires targets at heights of 60–130 m, which results in complex and variable image backgrounds and highly variable target scales. Moreover, most of the targets are small, leading to a significant challenge in detection. The results presented in Table 5 demonstrate that YOLO-ViT has significant advantages in detecting infrared vehicle targets for UAVs. Specifically, when the input resolution is 640 × 640, YOLO-ViT improves AP by 0.5%, 1.0%, and 1.1% for pedestrians, vehicles, and bicycles, respectively, compared to YOLOv5s with the same F1-score. However, the model complexity was 17.2% and 62.5% lower than YOLOv5m and YOLOv5l, respectively. Although the F1-score is lower than YOLOv5m and YOLOv5l by 0.3% and 0.6%, respectively, the model achieves better mAP. Additionally, when compared to the current state-of-the-art algorithm YOLOv8s, YOLO-ViT improves the F1-score by 0.3% and mAP by 1.0%. Notably, the detection accuracy AP for vehicles is improved by 1.8%, demonstrating a significant improvement in detection accuracy. The experimental outcomes described above demonstrate that the YOLO-ViT model, as proposed in this research paper, can attain superior detection accuracy and precision in identifying infrared vehicle targets, regardless of their overhead angles and sizes, even in the presence of intricate ground backgrounds. Furthermore, this model exhibits varying degrees of enhancement in detecting smaller targets, such as pedestrians and bicycles, highlighting its practical significance.

Table 5. Performance comparison with advanced algorithms.

Model	Size	Parameters	F1 (%)	AP_{Person} (%)	$AP_{Vehicle}$ (%)	$AP_{Bicycle}$ (%)	mAP (%)
YOLOv5s	640	7.0 M	90.6	92.8	97.1	91.0	93.7
YOLO5m	640	20.9 M	90.8	92.2	96.6	90.9	93.2
YOLO5l	640	46.1 M	91.2	93.2	96.9	90.6	93.6
YOLO7s	640	36.5 M	89.3	92.1	97.6	91.2	93.6
YOLO8s	640	11.2 M	90.3	92.6	96.3	91.5	93.5
YOLO-ViT	**640**	**17.3 M**	**90.6**	**93.3**	**98.1**	**92.1**	**94.5**

The results of different algorithms shown in Figure 7 reveal the significant improvement achieved by the YOLO-ViT approach for infrared target detection in UAVs. The YOLO-ViT method has demonstrated its ability to increase detection accuracy and decrease

the number of missed targets by learning feature information of infrared targets at various scales. Despite its proficiency, false alarms may still occur in dense target scenarios.

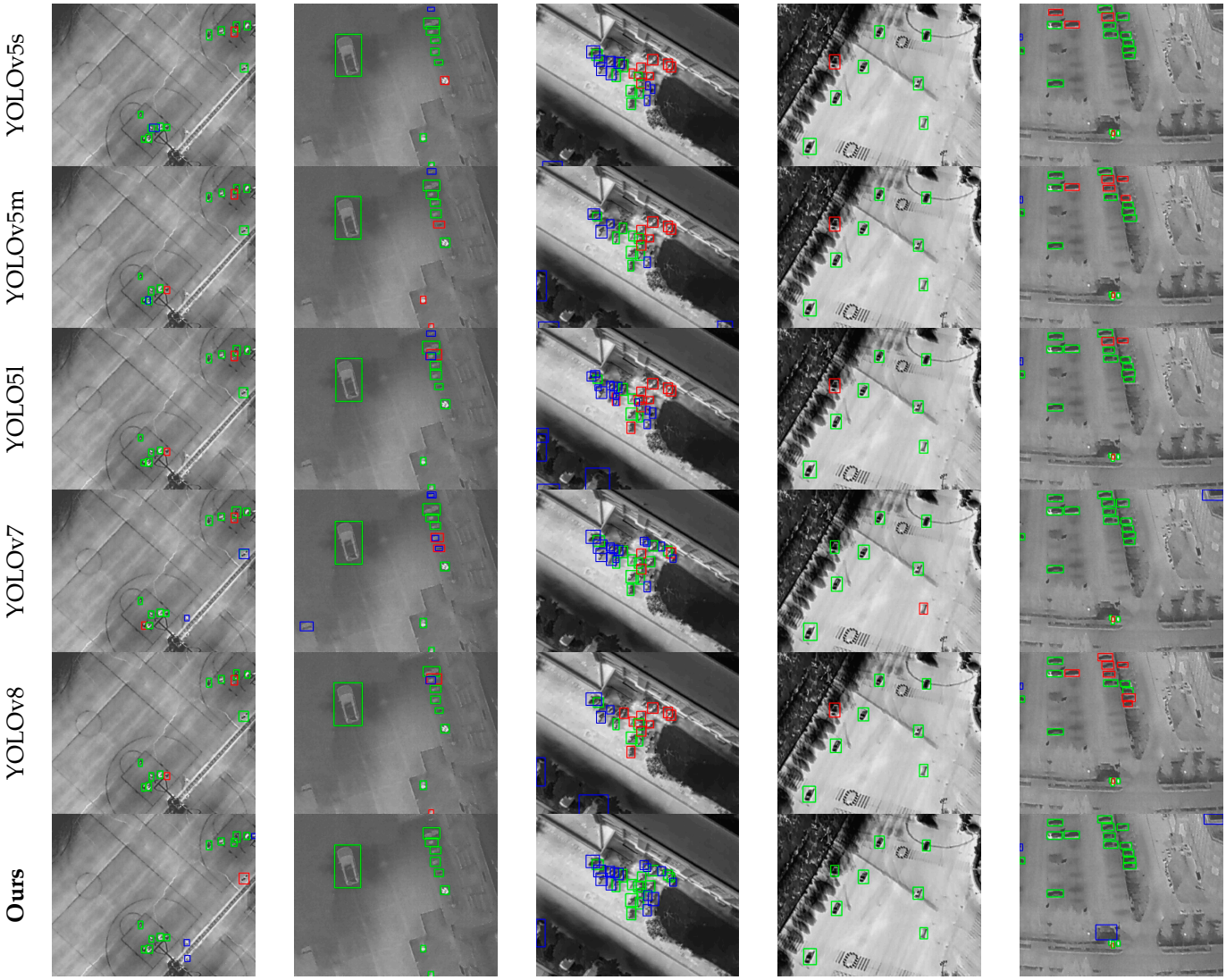

Figure 7. The outcomes of the various algorithms are presented through visualization, wherein the genuine positive (TP) instances are depicted by green boxes, while the false positive (FP) and false negative (FN) instances are represented by blue and red boxes, respectively. This visualization technique facilitates a clear understanding of the algorithmic results.

4.3. Ablation Studies and Analysis

In order to verify the effectiveness of the YOLO-ViT model proposed in this study, a series of ablation experiments were conducted on the HIT-UAV dataset to assess the impact of every individual module on YOLO-ViT. Specifically, the ablation experiment utilized an input image size of 640 × 640, a batch size of 16, and each network underwent training for 300 epochs. The outcomes of the experiments are presented in Table 6.

Table 6 presents that YOLO-ViT detection accuracy is notably improved under the same settings after optimizing the network structure as compared to the original YOLOv7. The upgraded MobileViT-based backbone network led to a 34.3% reduction in model complexity, a 1.5% decrease in mAP, and a 0.4% reduction in AP for vehicle detection accuracy. However, the AP for pedestrian and bicycle accuracy values for small targets were reduced by 1.8% and 1.3%, respectively. The improved MobileViT-based backbone network effectively reduced model complexity. The addition of the content-aware CARFE

upsampling operator resulted in a 0.2 M increase in parameters, a 0.1% increase in mAP, and AP values of 91.6%, 97.9%, and 91.4%, respectively. The F1-score also increased by 0.1%. Furthermore, the improved C3 module increased mAP by 0.6%, and there were 0.3%, 0.1%, and 1.2% accuracy improvements for different scales of targets. The modules mentioned above have the potential to enhance the detection accuracy of small targets by expanding the field of perception and optimizing the network structure in YOLOv7. Additionally, the K-means++ algorithm was utilized to pre-process the images and adjust the anchor frame size for small target detection, resulting in a more optimal anchor frame size. The performance of the modified algorithm was compared to the original algorithm, revealing a significant improvement in the F1-score of 1.2%, as well as an increase in AP accuracy, specifically by 1.3%, 0.5%, and 1.1%, respectively. Furthermore, there was an improvement of 1.1% in mAP, which substantially enhanced the accuracy of target detection.

Table 6. Ablation experiments based on the HIT-UAV UAV infrared dataset.

Yolov7	MobileViT	CARAFE	C3	K-Means ++	Parameters	F1 (%)	AP_{50}			mAP (%)	FPS
							Person	Vehicle	Bicycle		
√					36.2 M	89.3	92.1	97.6	91.2	93.6	54
√	√				23.8 M	87.5	90.3	97.2	88.9	92.1	39
√		√			36.4 M	89.4	91.6	97.9	91.4	93.7	51
√			√		29.9 M	89.9	92.4	97.7	92.4	94.2	60
√				√	36.5 M	91.5	93.4	98.1	92.6	94.7	54
√	√	√			26.7 M	89.2	90.0	97.4	90.3	92.6	37
√	√	√	√		17.3 M	89.6	90.6	97.6	90.5	92.9	40
√	√	√	√	√	**17.3 M**	**90.5**	**93.3**	**98.1**	**92.0**	**94.5**	**41**

Figure 8 demonstrates the superiority of the YOLO-ViT algorithm over the original algorithm in the domain of infrared vehicle target detection. The effectiveness of the enhanced modules in improving the accuracy of infrared target detection is also noteworthy. Therefore, the YOLO-ViT algorithm is a more appropriate solution for the task of infrared target detection in UAV scenarios.

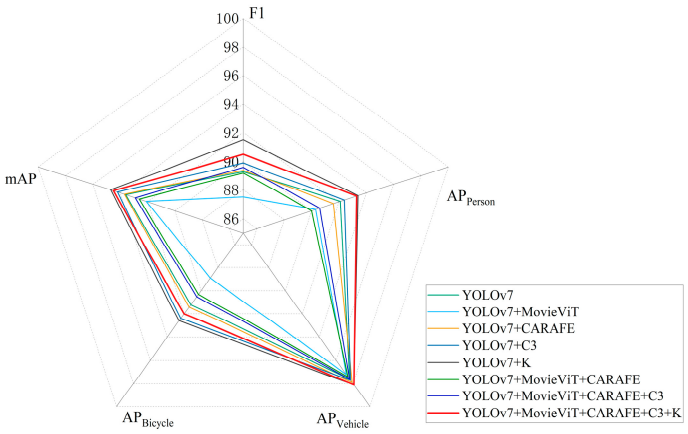

Figure 8. Ablation module performance comparison chart.

5. Conclusions

In this paper, we proposed a novel method, YOLO-ViT, for the detection of small and multi-scale infrared targets in the UAV scene. In the proposed method, a MobileViT-based backbone network is incorporated (blended) with the YOLOv7 model as the fundamental framework. The design of this network reduces the complexity of the model, enhances feature information fusion, and makes it more compatible with mobile devices. Moreover, the network's generalization is improved by incorporating the lightweight CARAFE upsampling and C3 modules, which increases the feature map's perceptual field while reducing

computational efforts and improving feature extraction and fusion of small infrared targets. To account for the specific characteristics of small targets in UAV aerial images, the K-Means++ clustering algorithm is used to preprocess the images and recalculate the anchor frame size, rendering it more suitable for IR target detection in UAV scenes. According to the results of the experiment, the proposed method exhibits a substantial reduction in both parameters and computational effort, amounting to 52.6% and 69.3%, respectively, when compared to the original YOLOv7. Furthermore, the mean average precision (mAP) has demonstrated an improvement of 0.9%, including a 0.5% enhancement in the accuracy of detecting vehicle targets. This reduction in the model complexity has rendered YOLO-ViT more precise and efficient in detecting infrared vehicle targets for UAVs, thereby making it a practical solution for UAV detection deployment.

Author Contributions: Conceptualization, X.Z.; validation, Y.X.; formal analysis, Y.X.; data curation, Y.X.; writing—review and editing, W.Z.; visualization, W.Z.; supervision, C.Z.; resources, Z.Z.; project administration, X.Z.; funding acquisition, X.Z. and Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 41404022, and the National Key Basic Research Strengthen Foundation of China, grant number 2021-JCJQ-JJ-0871.

Data Availability Statement: Data related to this study are available from the corresponding author upon reasonable request. The codes used during this study are available from the corresponding author upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wu, X.; Li, W.; Hong, D.; Tao, R.; Du, Q. Deep Learning for Unmanned Aerial Vehicle-Based Object Detection and Tracking: A survey. *Geosci. Remote Sens.* **2022**, *10*, 91–124. [CrossRef]
2. Qiu, Z.; Bai, H.; Chen, T. Special Vehicle Detection from UAV Perspective via YOLO-GNS Based Deep Learning Network. *Drones* **2023**, *7*, 117. [CrossRef]
3. Chen, Z.; Cao, L.; Wang, Q. YOLOv5-Based Vehicle Detection Method for High-Resolution UAV Images. *Mob. Inf. Syst.* **2022**, *2022*, 1828848. [CrossRef]
4. Ghasemi Darehnaei, Z.; Shokouhifar, M.; Yazdanjouei, H. SI-EDTL: Swarm intelligence ensemble deep transfer learning for multiple vehicle detection in UAV images. *Concurr. Comput. Pract. Exp.* **2021**, *34*, e6726. [CrossRef]
5. Du, Y. Multi-UAV Search and Rescue with Enhanced A∗ Algorithm Path Planning in 3D Environment. *Int. J. Aerosp. Eng.* **2023**, *2023*, 8614117. [CrossRef]
6. Choutri, K.; Mohand, L.; Dala, L. Design of search and rescue system using autonomous Multi-UAVs. *Intell. Decis. Technol.* **2021**, *14*, 553–564. [CrossRef]
7. Patel, T.; Guo, B.H.; van der Walt, J.D.; Zou, Y. Effective Motion Sensors and Deep Learning Techniques for Unmanned Ground Vehicle (UGV)-Based Automated Pavement Layer Change Detection in Road Construction. *Buildings* **2022**, *13*, 5. [CrossRef]
8. Cao, S.; Deng, J.; Luo, J.; Li, Z.; Hu, J.; Peng, Z. Local Convergence Index-Based Infrared Small Target Detection against Complex Scenes. *Remote Sens.* **2023**, *15*, 1464.
9. Zhang, R.; Newsam, S.; Shao, Z.; Huang, X.; Wang, J.; Li, D. Multi-scale adversarial network for vehicle detection in UAV imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *180*, 283–295. [CrossRef]
10. Srivastava, S.; Narayan, S.; Mittal, S. A Survey of Deep Learning Techniques for Vehicle Detection from UAV Images. *J. Syst. Archit.* **2021**, *117*, 102152. [CrossRef]
11. Bouguettaya, A.; Zarzour, H.; Kechida, A.; Taberkit, A. Vehicle Detection From UAV Imagery With Deep Learning: A Review. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 6047–6067. [CrossRef] [PubMed]
12. Gao, P.; Tian, T.; Zhao, T.; Li, L. GF-Detection: Fusion with GAN of Infrared and Visible Images for Vehicle Detection at Nighttime. *Remote Sens.* **2022**, *14*, 2771. [CrossRef]
13. Fan, Y.; Qiu, Q.; Hou, S.; Li, Y.; Xie, J.; Qin, M.; Chu, F. Application of Improved YOLOv5 in Aerial Photographing Infrared Vehicle Detection. *Electronics* **2022**, *11*, 2344. [CrossRef]
14. Yang, L.; Xie, T.; Liu, M.; Zhang, M.; Qi, S.; Yang, J. Infrared Small–Target Detection under a Complex Background Based on a Local Gradient Contrast Method. *Int. J. Appl. Math. Comput. Sci.* **2023**, *33*, 33–43.
15. Lin, T.; Maire, M.; Belongie, S. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 740–755.

16. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 23–28 June 2014. [CrossRef]
17. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015. [CrossRef]
18. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
19. Liu, S.; Ma, Z.; Chen, B. Remote Sensing Image Detection Based on FasterRCNN. In *Artificial Intelligence in China*; Springer: Berlin/Heidelberg, Germany, 2021. [CrossRef]
20. Wei, L.; Dragomir, A.; Dumitru, E.; Szegedy, C. *SSD: Single Shot MultiBox Detector*; Springer: Cham, Switzerland, 2016. [CrossRef]
21. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. [CrossRef]
22. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767. [CrossRef]
23. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696. [CrossRef]
24. Chen, W.; Baojun, Z.; Linbo, T.; Boya, Z. Small vehicles detection based on UAV. *J. Eng.* **2019**, *2019*, 7894–7897. [CrossRef]
25. Benjdira, B.; Khursheed, T.; Koubaa, A.; Ammar, A.; Ouni, K. Car Detection using Unmanned Aerial Vehicles: Comparison between Faster R-CNN and YOLOv3. In Proceedings of the 2019 1st International Conference on Unmanned Vehicle Systems-Oman (UVS), Muscat, Oman, 5–7 February 2019; pp. 1–6. [CrossRef]
26. Qiu, M.; Huang, L.; Tang, B.H. ASFF-YOLOv5: Multielement Detection Method for Road Traffic in UAV Images Based on Multiscale Feature Fusion. *Remote Sens.* **2022**, *14*, 3498. [CrossRef]
27. Liu, F.; Qian, Y.; Li, H.; Wang, Y. CAFFNet: Channel Attention and Feature Fusion Network for Multi-target Traffic Sign Detection. *Int. J. Pattern Recognit. Artif. Intell.* **2021**, *35*, 2152008. [CrossRef]
28. Liu, Y. Dense Multiscale Feature Fusion Pyramid Networks for Object Detection in UAV-Captured Images. *arXiv* **2020**, arXiv:2012.10643. [CrossRef]
29. Zhu, P.F.; Wen, L.; Bian, X.; Ling, H.; Hu, Q. Vision Meets Drones: A Challenge. *arXiv* **2018**, arXiv:1804.07437.
30. Sun, M.; Zhang, H.; Huang, Z.; Luo, Y. Road infrared target detection with I-YOLO. *IET Image Process.* **2021**, *16*, 92–101. [CrossRef]
31. Tang, T.; Zhou, S.; Deng, Z.; Zou, H.; Lei, L. Vehicle detection in aerial images based on region convolutional neural networks and hard negative example mining. *Sensors* **2017**, *17*, 336. [CrossRef] [PubMed]
32. Zhao, Q.; Liu, B.; Lyu, S.; Wang, C. TPH-YOLOv5++: Boosting Object Detection on Drone-Captured Scenarios with Cross-Layer Asymmetric Transformer. *Remote Sens.* **2023**, *15*, 1687. [CrossRef]
33. Zuo, Z.; Tong, X.; Wei, J.; Su, S.; Wu, P.; Guo, R.; Sun, B. AFFPN: Attention Fusion Feature Pyramid Network for Small Infrared Target Detection. *Remote Sens.* **2022**, *14*, 3412. [CrossRef]
34. Yao, S.; Zhu, Q.; Zhang, T.; Cui, W.; Yan, P. Infrared Image Small-Target Detection Based on Improved FCOS and Spatio-Temporal Features. *Electronics* **2022**, *11*, 933. [CrossRef]
35. Zhang, M.; Li, B.; Wang, T.; Bai, H. CHFNet: Curvature Half-Level Fusion Network for Single-Frame Infrared Small Target Detection. *Remote Sens.* **2023**, *15*, 1573. [CrossRef]
36. Li, S.; Li, Y.; Li, Y.; Li, M.; Xu, X. YOLO-FIRI: Improved YOLOv5 for Infrared Image Object Detection. *IEEE Access* **2021**, *9*, 141861–141875. [CrossRef]
37. Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Asymmetric Contextual Modulation for Infrared Small Target Detection. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 949–958. [CrossRef]
38. Zhang, M.; Zhang, R.; Yang, Y.; Bai, H.; Zhang, J.; Guo, J. ISNet: Shape Matters for Infrared Small Target Detection. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 867–876. [CrossRef]
39. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to Sequence Learning with Neural Networks. *Adv. Neural Inf. Process. Syst.* **2014**, 3104–3112. [CrossRef]
40. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929. [CrossRef]
41. Wang, W.; Xie, E.; Li, X.; Fan, D.; Song, K.; Liang, D.; Lu, T.; Shao, L. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; 2021. [CrossRef]
42. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.; EH Tay, F.; Feng, J.; Yan, S. Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet. *arXiv* **2021**, arXiv:2101.11986. [CrossRef]
43. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *Lecture Notes in Computer Science, Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Springer: Cham, Switzerland; Volume 12346. [CrossRef]

44. Liu, F.; Gao, C.; Chen, F.; Meng, D.; Zuo, W.; Gao, X. Infrared Small-Dim Target Detection with Transformer under Complex Backgrounds. *arXiv* **2021**, arXiv:2109.14379. [CrossRef]
45. Chen, G.; Wang, W.; Tan, S. IRSTFormer: A Hierarchical Vision Transformer for Infrared Small Target Detection. *Remote Sens.* **2022**, *14*, 3258. [CrossRef]
46. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. *arXiv* **2012**, arXiv:2012.12877.
47. Rao, Y.; Liu, Z.; Zhao, W.; Zhou, J.; Lu, J. Dynamic Spatial Sparsification for Efficient Vision Transformers and Convolutional Neural Networks. *arXiv* **2022**, arXiv:2207.01580. [CrossRef]
48. Suo, J.; Wang, T.; Zhang, X.; Chen, H.; Zhou, W.; Shi, W. HIT-UAV: A high-altitude infrared thermal dataset for Unmanned Aerial Vehicle-based object detection. *Sci. Data* **2023**, *10*, 227. [CrossRef]
49. Suo, J.; Wang, T.; Zhang, X.; Chen, H.; Zhou, W.; Shi, W. HIT-UAV: A High-altitude Infrared Thermal Dataset for Unmanned Aerial Vehicles. *arXiv* **2022**, arXiv:2204.03245.
50. Mehta, S.; Rastegari, M. MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer. *arXiv* **2021**, arXiv:2110.02178. [CrossRef]
51. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
52. Wang, J.; Chen, K.; Xu, R.; Liu, Z.; Loy, C.; Lin, D. CARAFE: Content-Aware ReAssembly of FEatures. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019. [CrossRef]
53. Arthur, D.; Vassilvitskii, S. K-Means++: The Advantages of Careful Seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, LA, USA, 7–9 January 2007.