drones

MDPI

Article

An All-Time Detection Algorithm for UAV Images in Urban Low Altitude

Yuzhuo Huang, Jingyi Qu *, Haoyu Wang and Jun Yang

Tianjin Key Lab of Advanced Signal Processing, Civil Aviation University of China, Tianjin 300300, China; 2021022111@cauc.edu.cn (Y.H.); 2022021020@cauc.edu.cn (H.W.); junyang@cauc.edu.cn (J.Y.)
* Correspondence: jyqu@cauc.edu.cn

Abstract: With the rapid development of urban air traffic, Unmanned Aerial Vehicles (UAVs) are gradually being widely used in cities. Since UAVs are prohibited over important places in Urban Air Mobility (UAM), such as government and airports, it is important to develop air–ground non-cooperative UAV surveillance for air security all day and night. In the paper, an all-time UAV detection algorithm based on visible images during the day and infrared images at night is proposed by our team. We construct a UAV dataset used in urban visible backgrounds (UAV–visible) and a UAV dataset used in urban infrared backgrounds (UAV–infrared). In the daytime, the visible images are less accurate for UAV detection in foggy environments; therefore, we incorporate a defogging algorithm with the detection network that can ensure the undistorted output of images for UAV detection based on the realization of defogging. At night, infrared images have the characteristics of a low-resolution, unclear object contour, and complex image background. We integrate the attention and the transformation of space feature maps into depth feature maps to detect small UAVs in images. The all-time detection algorithm is trained separately on these two datasets, which can achieve 96.3% and 94.7% mAP50 on the UAV–visible and UAV–infrared datasets and perform real-time object detection with an inference speed of 40.16 FPS and 28.57 FPS, respectively.

Keywords: UAM; UAV detection; all time; defogging detection; object enhancement

check for updates

Citation: Huang, Y.; Qu, J.; Wang, H.; Yang, J. An All-Time Detection Algorithm for UAV Images in Urban Low Altitude. *Drones* **2024**, *8*, 332. https://doi.org/10.3390/drones8070332

Academic Editors: Pablo Rodríguez-González, Giancarlo Rufino and Claudia Conte

Received: 10 May 2024
Revised: 14 July 2024
Accepted: 17 July 2024
Published: 18 July 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the gradual development of aviation technology, UAVs are extensively utilized in various domains of industry and daily activities, such as commerce and entertainment. On 1 January 2024, China formally implemented interim regulations on the Administration of Flight of Unmanned Aerial Vehicles issued by the State Council [1]. The document focuses on emphasizing that UAV flight activities should be safety-oriented to ensure orderly flight activities. In August 2022, the Civil Aviation Administration of China (CAAC) released the Civil Unmanned Aerial Vehicle Development Roadmap [2]. The document pointed out that the development direction of UAVs is urban high-density transportation, inter-city long-distance demand, and expanding to all-time operation capability.

The Office of the National Air Traffic Commission has clearly defined the flight altitude of urban low-flying drones, which are divided into areas such as airborne no-fly zones, airborne restricted zones, and temporary no-fly zones, taking into account the type of airspace restriction and the purpose for using them. Airborne no-fly zones refer to the airspace above nationally important political, economic, military, and other core vital targets, and no aircraft may fly into the airborne no-fly zones without approval. Temporary no-fly zones are temporary time limits set to restrict the entry of aircraft to ensure the safe and effective conduct of activities. Therefore, it is vital to ensure the safety of no-fly zones in cities by monitoring Unmanned Aerial Vehicle flight activities and using communication, navigation, and surveillance capabilities to ensure airspace safety.

There are several difficulties in the all-time detection of urban low-altitude UAVs in images as shown in Figure 1. As depicted in Figure 1a, the size of the UAVs in the images is

too small for detection; Figure 1b displays the issues with foggy weather; in Figure 1c, the urban background is relatively complex; and Figure 1d displays the issue of low resolution of infrared images.

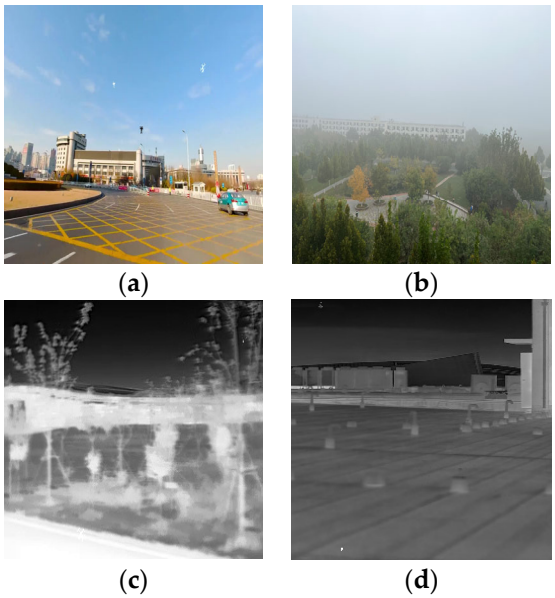


Figure 1. The challenges with all-time detection in the city: (**a**) small object; (**b**) foggy weather; (**c**) complex background; (**d**) low resolution of infrared images.

An all-time detection algorithm for UAV images in urban low altitudes is proposed, and our contributions can be outlined as follows:

(1) An algorithm for the real-time detection of urban low-altitude multi-scale UAVs in images is proposed by us. It is used for UAV detection tasks with different image characteristics during the daytime and nighttime. In the daytime, we propose the dehaze detection structure to solve the detection problem in foggy environments. At night, we propose a Squeeze and Excitation-backbone (SE-backbone) structure and a Space-to-Depth–Path Aggregation Network with the Feature Pyramid Network (SPD-PAFPN) structure to obtain effective information of deeper feature maps, which are used in UAV detection in low-resolution images;

(2) The two urban low-altitude UAV datasets based on an adaptive approach are constructed. They are used for the training of the models under daytime and nighttime conditions. The datasets were employed to resolve the lack of data in the UAV object detection research in urban visible and infrared conditions.

We utilize deep learning technology to perform all-time detection on UAV images with complex urban low-altitude backgrounds. For all-time surveillance of non-cooperative targets, visible light images are employed during the day and infrared images at night, ensuring all-weather UAV monitoring. This approach, in conjunction with existing active surveillance equipment such as radars, jointly safeguards urban air traffic.

2. Related Work

In the all-time object detection task, it is important to choose an appropriate detection method for the different characteristics of daytime and nighttime images. The predominance of small targets leads to limited feature information available for UAV objects. In application, UAV detection requires better detection accuracy and real-time performance.

Since the successful application of the AlexNet [3] convolutional neural network in vision tasks, object detection with deep learning methods has been widely popularized, such as ResNet [4], DenseNet [5], MobileNet [6], ShuffleNet [7], and the YOLO series of algorithms [8–11]. Based on the wide application of convolutional networks, many

researchers have been dedicated to improving the existing model structures, such as CrossConv [12], TransFormer [13], GnConv [14], SwiftFormer [15], and CoordConv [16]. These improvements are all in search of better detection results.

Object detection in visible images under sunny weather conditions often employs methods such as the enhancement of target feature information. Yu et al. [17] proposed a better scale conversion algorithm for detecting tiny pedestrians in a city. Ye et al. [18] proposed a CT-Net neural network for balancing the accuracy, model size, and real-time detection of small objects, which achieved good results. Zeng et al. [19] proposed a feature fusion method for small object detection and enhanced the semantic information of small objects. Liu et al. [20] proposed a DO-PAN structure for small objects with little feature information, which predicts the feature maps of different depths of small object information on the same scale and enhances the capability of the model for small object detection. Minaeian et al. [21] presented a new object detection and localization visual system for UAVs. Zhang et al. [22] proposed a model to solve UAV aerial image detection based on an improved YOLOv5. Zhang [23] designed and proposed an algorithm based on the YOLO to solve the problem of small target detection in UAV images.

During inclement weather conditions, such as rain, snow, fog, haze, and other weather environments, it is more difficult to detect small objects. In complex urban background conditions, Huai et al. [24] proposed a multi-object detection algorithm for urban automobiles incorporating historical trajectories. The method can realize multi-object high-precision detection in complex urban environments with object occlusion, light changes, and cloudy or rainy weather. For the detection task in low light and hazy weather, Liu et al. [25] proposed the IA-YOLO method based on YOLOv3. The method incorporates an adaptive filter to dehaze the image and is applied to pedestrian detection with improved accuracy. Sanket [26] proposed a GDIP gating mechanism based on IA-YOLO, which allows multiple microprocessor image processing modules to be concurrently relative-weighted to enhance object detection under different conditions. Mazin et al. [27] proposed a multi-scale domain adaptive MS-DAYOLO algorithm. This method uses domain adaptive paths and classifiers at different scales of the YOLOv4 to generate features, which significantly improves the performance on foggy days. However, the practical requirement is to achieve the defogging effect on foggy images without distortion on clear weather images. Therefore, we propose a combination of a defogging network GridеDehazeNet [28] based on the YOLOv7 detection network. The proposed defogging detection structure implements a multi-scale estimation in the backbone module. The structure does not rely on traditional atmospheric scattering models and can provide more flexible and objected image defogging. It ensures that the images before and after the input and output are not distorted, which is beneficial for further UAV object detection tasks.

Infrared sensors at night can realize continuous passive detection by receiving infrared radiation and converting it into infrared images, which has strong practical application value. However, there are issues with the low resolution of infrared images and it is subject to a wide variety of noise interference during detection. The flight background is complex under the interference of urban buildings and natural weather conditions, and small infrared objects are easily submerged. These reasons make it difficult to accurately detect UAVs.

In infrared image object detection algorithms, researchers resolve the small object detection problem by obtaining more feature information. Raja et al. [29] proposed space-to-depth convolution (SPD-Conv), a new convolution module for low-resolution images. This module is used instead of stepwise convolution and a pooling layer to obtain more fine-grained information, which is applied to the YOLOv5 model to improve the accuracy of infrared image detection. Zhao et al. [30] proposed an infrared object detection model based on YOLOv8 for UAV images to detect the complex background targets on the ground. It is difficult to detect small objects because there is a lack of feature information. To eliminate the problem, we propose an improved structure for low-resolution infrared images. The SE attention is integrated in the max pooling convolution (MPConv) structure of the YOLOv7

backbone network [31]. It is easier to acquire more object feature information of low-resolution images before down-sampling. Due to the long distance of ground and air surveillance, most of the UAVs under actual surveillance are small objects; therefore, a spatial-depth transformation module SPD-Conv is used for the small object detection head of the detection network. This module splits the feature map into multiple sub-feature maps and connects the sub-feature along the channel dimension. By acquiring more feature maps, this focuses on more object information. Our method has shown improvement for object detection in low-resolution images.

Existing detection methods fall short in offering all-time UAV detection specifically tailored for urban low-altitude environments. A single model's performance deteriorates when attempting to handle both day and night conditions simultaneously. In this paper, an algorithm for all-time detection of urban UAVs in images is proposed by us. The method is able to solve the problem of low accuracy on foggy days. It also solves the problem of small UAV detection in low-resolution infrared images in the form of acquiring more feature information. In addition, two publicly available urban low-altitude UAV datasets are provided. The method in this paper is trained, tested, and achieved excellent detection performance on the two datasets.

3. Data Setting

Urban UAVs fly in complex environments and the flight background changes all the time. The complex background creates difficulties for UAV object detection. On the one hand, the different observation distances of UAVs lead to different sizes of the UAVs in the image. The change in object scale also brings challenges to the detection algorithm. On the other hand, the learning of neural network algorithms requires many datasets. The quality of the dataset largely reflects the detection results of the final model as well. However, for urban low-altitude UAVs, there are basically no datasets publicly available. In summary, two urban low-altitude complex background UAV datasets based on visible and infrared images are constructed.

Due to the fact that there are few low-altitude flying UAVs in the city at this stage, it is not enough to construct the dataset for training. Therefore, this paper proposes data synthesis as the solution idea to construct the dataset. According to the solution idea, this paper firstly constructs a UAV gallery. The gallery comes from the actual UAVs and the UAV pictures are collected from research. It contains drones with different flight positions, different brands and models, and different colors. Since the drone pictures have different backgrounds, it is difficult to synthesize the data in the next step. Therefore, the image preprocessing technique is utilized to remove the complex background interference. We obtained the UAV images with a transparent background, as shown in Figure 2. Figure 2a displays UAVs with cameras, Figure 2b shows single-wing UAVs, and Figure 2c displays multi-rotor UAVs. Eventually, all the UAV images with a transparent background were constructed to form a UAV gallery.

On the basis of the UAV gallery, our team also constructed a visible and infrared city street view gallery. As shown in Figure 3, the visible images use real-life visible light pictures of complex urban backgrounds. Figure 3a displays street view images with sky as the main background; Figure 3b shows street view images of building backgrounds; and Figure 3c displays street view images in foggy weather. As shown in Figure 4, the infrared images are adopted from the publicly available Arrow Photonics database dataset of urban nighttime infrared images. The urban street view images contain different backgrounds such as high-rise buildings, sky, trees, and billboards. It conforms to the background environment characteristics of the city's low-altitude flight, thus constructing the visible and infrared view gallery.

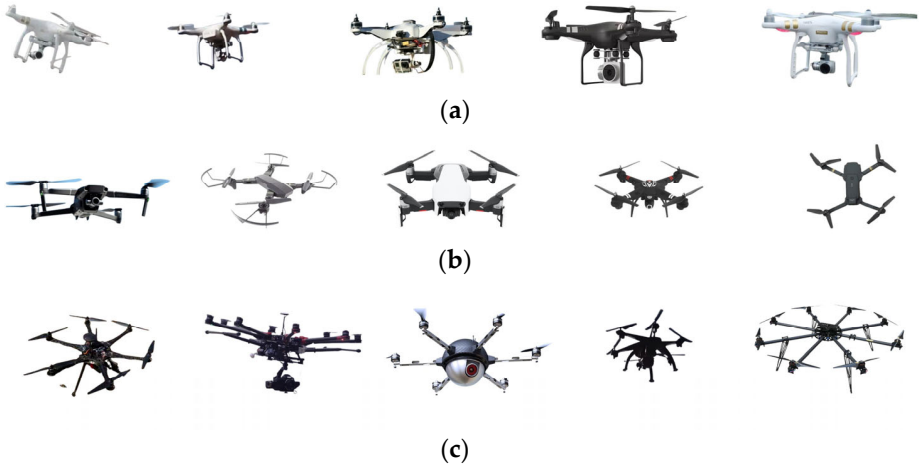


Figure 2. UAV images: (**a**) UAVs with cameras; (**b**) single-wing UAVs; (**c**) multi-rotor UAVs.

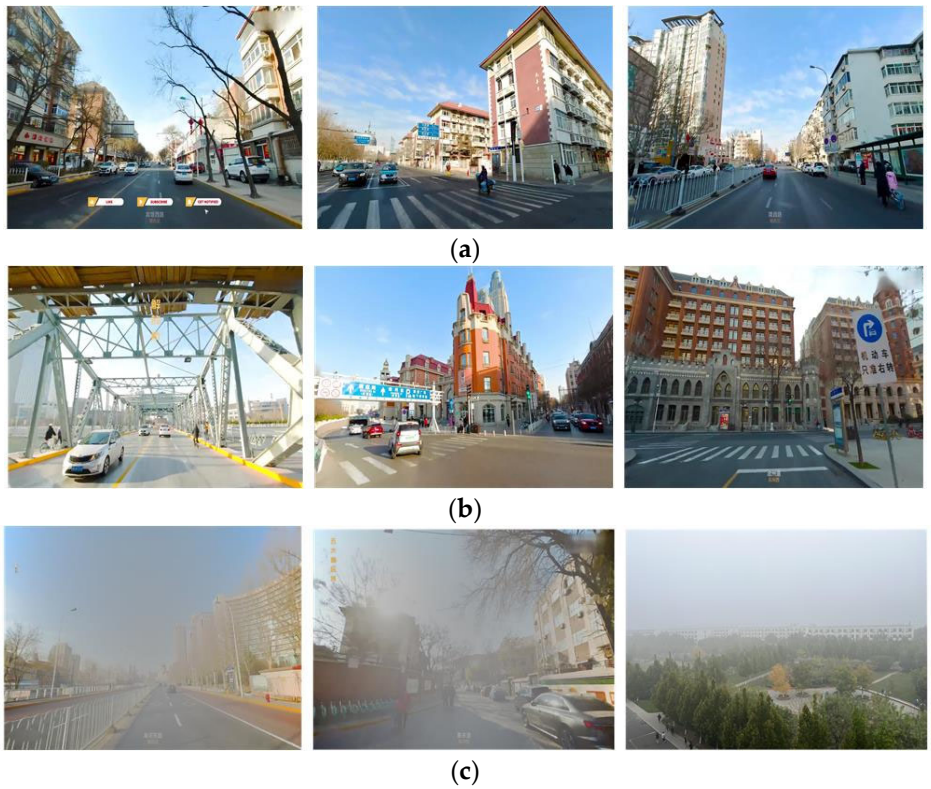


Figure 3. Urban visible images: (**a**) street view images with sky as the main background; (**b**) street view images of building backgrounds; (**c**) street view images in foggy weather.

On the basis of the UAV and street view image library, the method of data synthesis is utilized. We scaled the obtained UAV pictures according to a certain ratio. We select the middle- and upper-part of the street view images for data synthesis at random locations, thus obtaining the urban low-altitude UAV flight dataset. The visible dataset contains clear weather as well as fogged UAV flight images. In order to make the constructed dataset more realistic than just synthesized UAV data, the visible dataset was constructed by including real foggy weather images and images of UAV flights taken with the airport as the background. The dataset is named the UAV dataset used in an urban visible background (UAV–visible). It contains a total of 1380 foggy images with different concentrations. The total number of clear weather images is 3210, with a total of 4590 images containing 4590 UAV objects. The training set is 67.2%, the validation set is 16.8%, and the test set is 16%,

and all the images are 1920 pixels × 1080 pixels in size. As shown in Table 1, the objects are differentiated in terms of their proportion of the whole image size according to (1) [32], the *px* indicates the number of pixels in the object area. The dataset contains 3070 small objects, accounting for 72.23%; 962 medium objects, accounting for 22.64%; and 218 large objects, accounting for 5.13%. It satisfies the status quo that all kinds of scale objects are available when urban UAVs are flying.

$$\text{Object } Area = \frac{px}{\sum_{i=1}^{n} px_i} \times 100\%. \quad (1)$$



Figure 4. Urban infrared images.

Table 1. UAV object size distribution on the UAV–visible dataset.

Object	Object Area	Quantity	Quantity Share	Classification
UAV	<0.12%	3070	72.23%	Small
	<0.48%	962	22.64%	Medium
	≥1.9%	218	5.13%	Big

According to the principle of constructing UAV–visible. In this paper, we also screened the infrared images of UAVs in the background of airports that were actually captured using infrared equipment. The urban low-altitude UAV dataset based on infrared images is thus constructed, and the dataset is named the UAV dataset used in urban infrared background (UAV–infrared). The dataset comprises a total of 3400 images containing 6800 UAV objects, and the image sizes are all 256 × 256 pixels. As shown in Table 2, the training set in the experiment is 67.2%, the validation set is 16.8%, and the test set is 16%. According to the division standard of relative size, there are 4180 objects belonging to small objects in this dataset, accounting for 61.47%; 2390 objects belonging to medium objects, accounting for 35.15%; and 230 objects belonging to large objects, accounting for 3.38%. It satisfies the status quo of having objects of various proportions when flying urban drones.

Table 2. UAV object size distribution on the UAV–infrared dataset.

Object	Object Area	Quantity	Quantity Share	Classification
UAV	<0.12%	4180	61.47%	Small
	<0.48%	2390	35.15%	Medium
	≥1.9%	230	3.38%	Big

4. Methodology

In this section, an all-time UAV detection method based on visible images during the day and infrared images at night is proposed by our team. In Figure 5, the model consists of three parts: input image, model, and output image. First, the input images include visible and infrared images, which are important to the model for detection. Second, the detection model consists of three parts: mode selection, visible detection, and infrared detection. Finally, the output image contains the position and confidence information of the UAV objects.

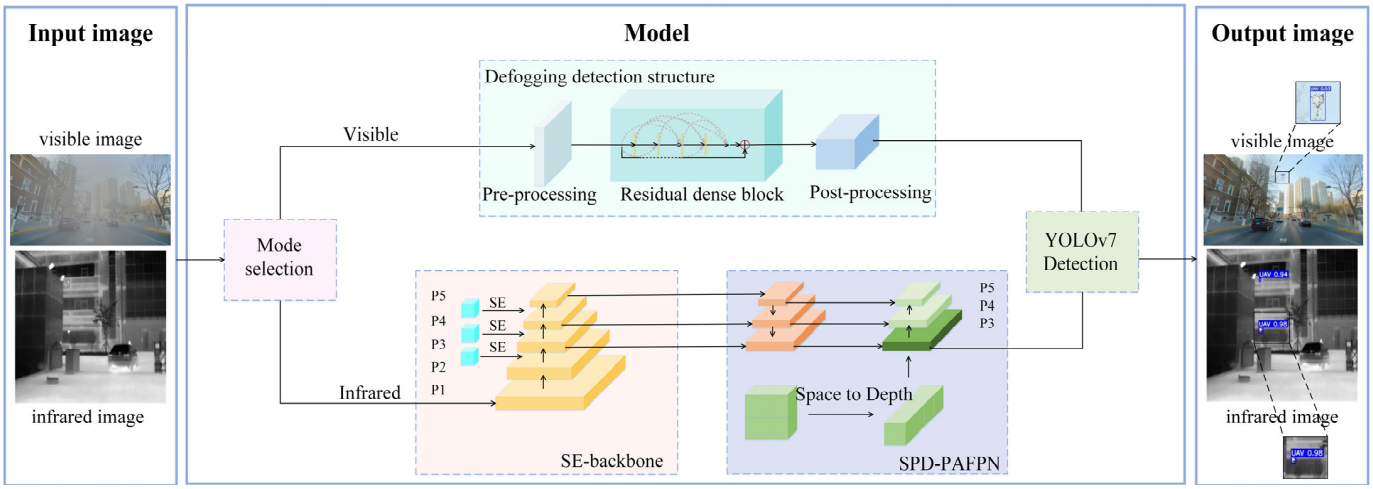


Figure 5. All-time detection network model.

4.1. Mode Selection

In the model, the input image is first passed through a mode selection module that is carried out. According to the different characteristics of the pictures, the corresponding visible and infrared modes are selected. Since the picture features of visible and infrared pictures are quite different, this paper chooses to analyze the mean and variance features of the pictures for judgment.

The mean value of an image is the average gray value of all pixel points in a gray-scale image. The mean value of the image can reflect the brightness of the whole image. If the mean value of the image is larger, the image is brighter. If the mean value of the image is smaller, the brightness of the image is lower. The variance of an image is the average of the squared values of the difference between the gray value of each pixel in the image and the mean gray value of that image. The variance of an image reflects the contrast of the entire image. If the variance of the images is higher, the images will have higher contrast. If the variance of the images is smaller, the contrast of the images is lower. The higher the contrast of an image, the sharper, clearer, and more striking it appears.

The lower the contrast of the image, the less sharp the image is. In all-time detection, the image undergoes changes in clarity as well as gray level over time. In this paper, the product of the mean and variance of an image is used as the basis for different mode selection. If the product of the mean value and variance is greater than the threshold, then it is in visible mode, and if it is less than the threshold, then it is in infrared mode. For the selection of the threshold value, this paper adopts the threshold value of 1000 obtained after

experimental analysis [33]. In addition, we utilize time as an auxiliary means to determine whether it is daytime or nighttime. This is aimed at better leveraging the detection models suited for different conditions.

4.2. Defogging Detection Structure

To solve the problem of UAV fog detection in urban low altitudes, we combine the GrideDehazeNet [28] network with YOLOv7 [10] (You Only Look Once, YOLO) for defogging detection of UAVs in the city. The model for defogging detection consists of four modules: preprocessing, residual dense block, post-processing, and detection.

4.2.1. Preprocessing

The preprocessing module comprises convolutional layers and residual dense blocks. It generates 16 feature maps from the given foggy image, which will be referred to as learned inputs. It is trainable and undergoes optimization throughout the network's training process, enabling the backbone network to perform dehazing more effectively.

The preprocessing methods generated in this manner offer better diversity and more object features than manually generated methods. The module generates multiple foggy images with the same features so that the problem can be transformed into multiple images. It highlights aspects of the image and enables a more pronounced focus on relevant feature information.

4.2.2. Residual Dense Block

A residual dense block (RDB) arrangement forms a grid structure with three rows and six columns. The composition of five RDB blocks per row of the network allows the number of feature maps to remain constant. In each sampling block, the size of the feature maps is varied with a scale of two, and the number of feature maps is varied accordingly. The internal structure of the RDB blocks is shown in Figure 6. An RDB composed of five convolutional layers can increase the number of feature maps. A convolution with a stride of 2 is used to uniformly fuse all feature maps, at the same time, fusing the output of the previous step with the up-sampled feature.

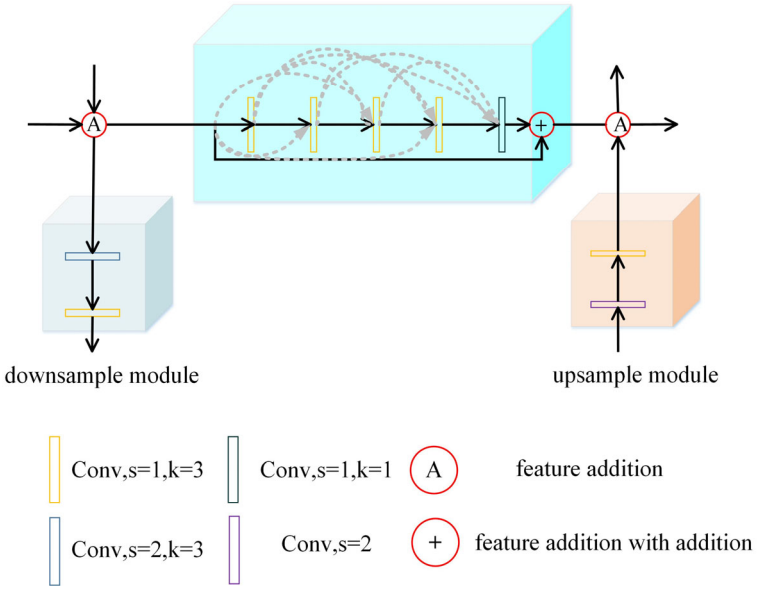


Figure 6. RDB model structure.

In the feature fusion stage, an attention is introduced to regulate the weights of different scales of information. The output of the horizontal block RDB and the output of the vertical RDB block are fused, and the fusion mechanism is shown in (2) where $\widetilde{F}^{i}$

represents the result after fusion of channel i; a_r^i represents the i channel of the current fused horizontal input; $\widetilde{F}^{\mathrm{i}}$ represents the fusion weight of the i channel of the course fused horizon F_c^i input; F_c^i represents the i channel of the current fused vertical input; a_c^i represents the fusion weight of the i channel of the current fused vertical input.

$$\widetilde{F}^{\mathrm{i}} = a_r^i F_r^i + a_c^i F_c^i. \tag{2}$$

4.2.3. Post-Processing

The post-processing module is mainly used to perform further operations such as color correction, contrast enhancement, and color balancing on the defogging results to further improve the visual effect of the image. By enhancing the edges and details of the image, the defogging results can be made to look clearer and more realistic in order to reduce the noise level in the image. As fogged images may have color bias or color distortion, the post-processing module introduces color correction methods to improve the color accuracy and reproduction of the image. The defogging results generated by the network are further optimized to improve the visual quality so that the defogged images are more in line with the perception of the human eye.

The post-processing module achieves better results compared to manual processing. After the defogging network outputs a fog-free image result, this fog-free image is used as an input to the object detection.

4.2.4. Detection

The object detection network is selected as the YOLOv7 network. The detection speed of this model meets the requirement of real-time detection and localizes as well as classifies the large, medium, and small UAVs in the detection head. The defogging detection model does not depend on the traditional atmospheric scattering model in the defogging stage. This approach solves the problem of picture distortion before and after defogging. The deep learning-based approach can provide clearer pictures, which can make picture defogging more effective. It is more conducive to carry out the next UAV detection task. We take the image output from GrideDehazeNet as the input to YOLOv7. The image is scaled with a uniform size to extract the feature information of objects of different sizes. The final output detection result contains the category of the object as well as the confidence level.

4.3. SE-Backbone Structure

It is more difficult to detect drones in infrared images at night in urban low altitudes. Infrared images are characterized by low resolution, complex backgrounds, and small objects. The more layers a convolutional network has, the more feature information it can obtain. At the same time, it will lower the resolution of the feature map, which is unfavorable for detecting small-object UAVs. Therefore, the SE-backbone structure based on MPConv in backbone networks is proposed by us. We fuse the SE attention mechanism in several different locations of the MPConv before the extracted feature maps are subjected to down-sampling operations. It can be called the MPConv-SE in the paper. A comparison of the module structure before and after improvement is shown in Figure 7. Meanwhile, we constructed MPConv-SA [34] and MPConv-GAM [35] to be compared in the experiments.

MPConv-SE improves on the original top and bottom two-branch structure. The first branch consists of max pooling and a 1×1 convolution that changes the number of channels after a max-pooling operation. In the second branch, the first 1×1 convolution of the original structure is replaced by a SE module. The rest of the structure remains unchanged, and finally, the output of the top and bottom branches are spliced together.

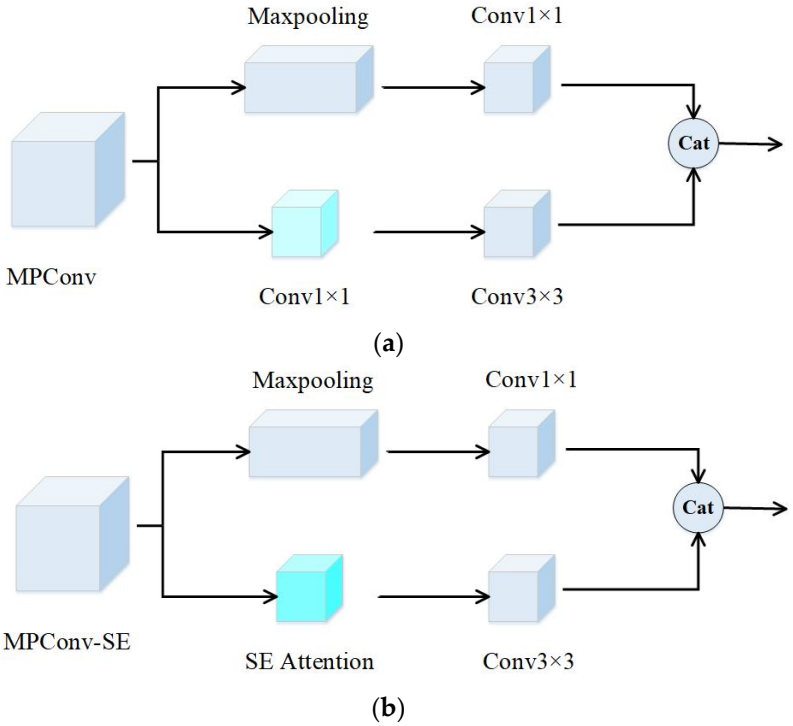


Figure 7. Comparison of MPConv and MPConv-SE structures: (**a**) MPConv structure; (**b**) MPConv-SE structure.

The structure of the SE [31] channel attention introduced in the second branch down-sampling in this paper is shown in Figure 8. The SE module first performs the Squeeze operation as in (3) on the convolution to obtain the feature map. It encodes and compresses all the spatial features on a channel into one global feature. Then the global features at the channel level are obtained using global level pooling. The Excitation operation as in (4) is then performed on top of this to output features by multiplying the obtained weights of different channels with the original feature map. With such a weight matrix, the importance of the channels is measured by assigning different weights to different positions of the input image in the channel dimension. The problem of loss due to different importance of each channel in the down-sampling is solved.

$$F_{\mathrm{sq}}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j), \tag{3}$$

$$F_{\mathrm{ex}}(z, W) = \sigma(g(z, W)) = \sigma(W_2 RELU(W_1 z)). \tag{4}$$

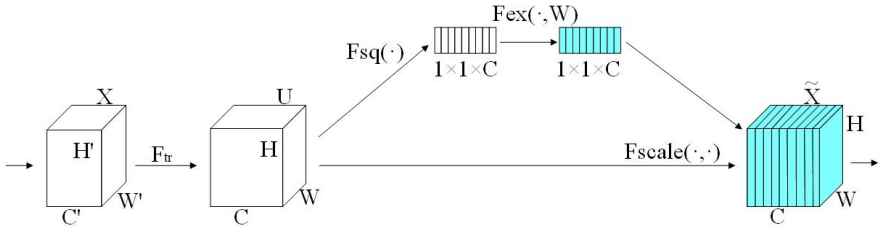


Figure 8. Structure of the SE channel attention.

Since the second 3 × 3 convolutional structure of the second branch acts as down-sampling, it can shrink the feature map to reduce the dimension. Therefore, the convolution in front of it is replaced with the SE attention mechanism module. In order to obtain more

feature information before down-sampling, the feature map after dimensionality reduction retains more effective information and reduces the leakage and false detection of UAVs.

4.4. SPD-PAFPN

Urban infrared images have a low resolution, complex environment, and a small number of feature maps. These characteristics make it difficult to train and obtain a high-performance UAV detection model. In order to obtain more information about the UAV features in low-resolution images, the SPD-Conv [29] is added to the small object detection layer of PAFPN. The structure of the model is shown in Figure 9. The module consists of a space-to-depth layer and a non-spanning convolutional layer. This combination reduces the size of the spatial dimension without losing information. It also retains the information in the channel, which helps to improve the performance of CNN for detecting small object UAVs in low-resolution images.

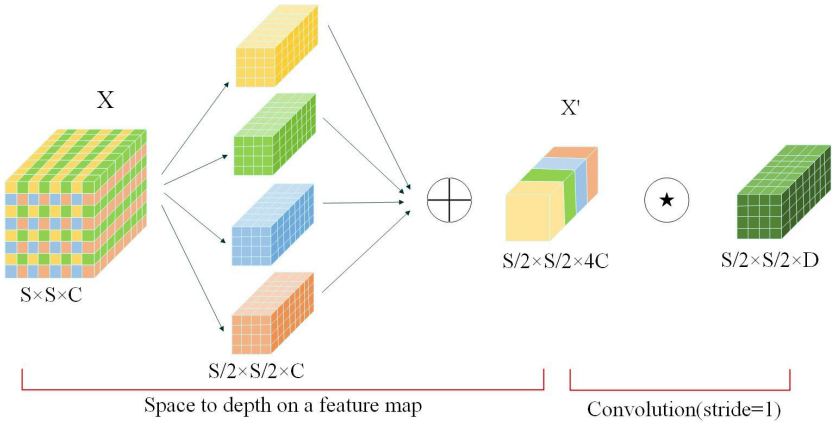


Figure 9. Structure of the space-to-depth model. The star symbol represents convolution.

We use this structure to reduce the dimensionality of input feature maps. The input feature map X is of size S × S × C. When the scale factor scale = 2, the original feature map X is sliced into four sub-feature maps with feature identification information according to a scale of 2. The size of the obtained feature map is S/2 × S/2 × C. The sub-feature maps are then connected along the channel dimension, which is used to obtain a feature map X′ of size S/2 × S/2 × 4C. This method avoids excessive feature information loss.

The convolutional layer with the stride of 1 is connected after acquiring the features. Unlike stepwise convolution, non-stepwise convolution does not move over the feature map. Instead, it performs a convolution operation on each pixel feature map. Such an approach helps to reduce the problem of oversampling that may occur in the SPD layer and retains more fine-grained information. The non-stepwise convolutional layer has 4C filters, $4C < (scaleC)^2$. The feature map X′ is further transformed so that as much discriminative feature information as possible can be retained to facilitate the detection task. The specific calculation formula is shown in (5), where s is the independent variable and takes a natural number greater than or equal to 1.

$$f_{s-1,s-1} = X[s-1:S:s, s-1:S:s]. \tag{5}$$

As for solving the problem of UAV detection in low-resolution images, the SPD-Conv is added in the PAFPN feature fusion stage. This can obtain more effective feature information in low-resolution images for detecting UAVs with small objects. In addition, we can use the P4 and P5 layers to detect medium or large objects.

5. Experiments and Analysis

Our proposed all-time detection algorithm has good performance for the detection of UAV objects with multi-scale sizes. We validate the feasibility of our method through

comparative experiments and ablation experiments. In this section, we cover the model experimental environment and parameter settings in turn. The evaluation indexes of this experiment are listed in the paper and the results are analyzed.

5.1. Experimental Equipment and Hyperparameter Settings

The computer configurations used for the experimental model training and inference speed test in this paper are as follows. The GPU accelerated graphics card is Tesla T4. The operating system is a 64-bit Ubuntu 16.04.7 system. The deep learning development environment employs Pytorch 1.8.1. The data are divided into three parts: training set, validation set, and test set, according to the ratio of 7:2:1. The parameter configuration of the network model is shown in Table 3.

Table 3. Table of configuration of experimental parameters.

Parameters	Values
Mosaic	1.0
Optimizer	SGD
Learning Rate	0.01
Weight Decay	0.0005
Loss Function	CIoU
Training Batch	8

5.2. Evaluation Metrics

The evaluation metrics used in this experiment are mean accuracy (mAP), probability of detection (P), recall (R), F1-score, and model prediction speed in the unit of Frames Per Second (FPS). The mAP is the average accuracy of all categories and is calculated by the formula as in (6). mAP50 is the prediction value of AP when the intersection of prediction frame and truth frame is thresholded to be 0.5. The larger the value is, the proof that the model's overall detection performance is better. The average accuracy described in this paper is the mAP50. The P is the probability of predicting all the samples to be positive samples, calculated as in (7). The R is the probability of the actual positive samples to be correctly predicted by the model, calculated as in (8). The F1-score indicates the tuning level of the P and R for different models, calculated as in (9). FPS reflects the refresh rate of the model's inference speed and is calculated as in (10).

$$mAP = \frac{1}{c}\sum_{i=1}^{c} AP_i, \tag{6}$$

where c is the number of all categories; APi is the area of the i recall on the P-R curve with respect to the corresponding precision curve.

$$P = \frac{TP}{TP + FP}, \tag{7}$$

$$R = \frac{TP}{TP + FN}, \tag{8}$$

The True Positive (TP) is the number of correct results, the False Positive (FP) is the number of incorrect results, and the False Negative (FN) is the number of correctly detected objects that are missed.

$$F1 - score = \frac{TP}{TP + FP}, \tag{9}$$

The F1-score is the reconciled average of correctness and recall. The F1-score value ranges from 0 to 1, with values closer to 1 indicating better model performance. When both

precision and recall are high, the F1-score value is also higher. This indicates that the model is able to achieve a more balanced performance in the detection task.

$$FPS = \frac{1}{\text{inf}} \tag{10}$$

The inf represents the model inference time. The time required for the model to make predictions for a single input image.

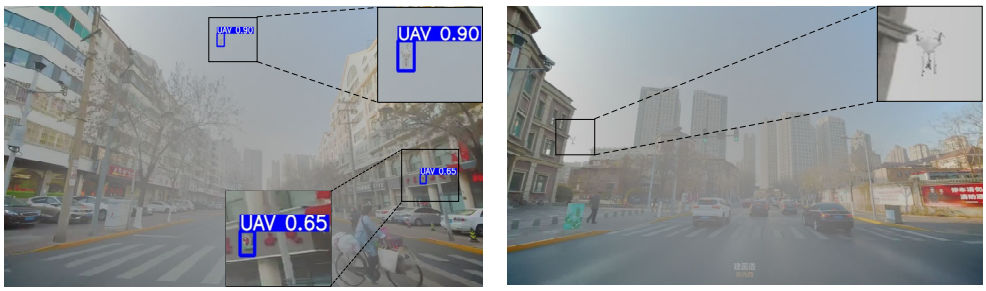
5.3. Experimental Analysis of Visible Images

Table 4 shows the results of detecting UAVs on the UAV–visible dataset using different object detection models. From the results, the YOLOv7 selected in this paper is outstanding. The detection on the test set can achieve 95.6% accuracy with a real-time performance of 39.22 FPS, which meets real-time detection requirements.

Table 4. Comparison of detection results of different models in visibility.

Model	Precision(%)	Recall(%)	mAP50(%)
Faster-RCNN [36]	75.3	70.2	74.4
SSD [37]	75.2	68.8	72.2
YOLOv5s [8]	86.5	81.9	85.7
YOLOv6 [9]	92.1	84.4	92.4
YOLOv8n [11]	94.5	89.0	93.5
YOLOv7 [10]	96.7	92.1	95.6
IA-YOLO [26]	94.5	90.0	93.8
Ours (visible)	**97.7**	**93.2**	**96.3**

Figure 10 shows our analysis of the results of the detection. It can be found that there are still misdetections and omissions in the output results. The results can be improved in detection accuracy. From the results, we can see that some of the detection results under foggy conditions mistakenly detect the trees obscured by fog as drones. There are also foggy conditions that result in drones not being accurately detected. To summarize, we can see that foggy weather has a greater impact on UAV object detection. Foggy days lead to lower visibility of the field of view, resulting in poor accuracy in UAV detection. Therefore, the defogging processing can be used to resolve the problem. The input image is first processed to defog before the object detection task. The original clarity of the image is restored as much as possible. It can reduce the detection task phenomenon of leakage and misdetection. This method solves the problem of urban low-altitude complex background UAV detection in foggy weather conditions.



(**a**) (**b**)

Figure 10. Analysis of UAV testing results in foggy weather: (**a**) misdetection; (**b**) missed detection.

We utilize the GrideDehazeNet to defog the UAV–visible dataset to obtain clear and fog-free images, and then input this result into YOLOv7 for object detection. In this paper, the IA-YOLO [26]—which has outstanding detection effect in low light and foggy days—is selected for comparative experimental analysis. The output results after defogging are

shown in Figure 11. The first row is the original image of the dataset, and the second row is the IA-YOLO detection result graph. The third row is the detection output result graph of our method.



Figure 11. Comparison of detection results of the IA-YOLO and ours.

By selecting the same picture for detection and comparison, there is no distortion in the output image of our proposed method of first defogging and then detecting, while the output of the IA-YOLO method produces serious distortion. When applied to actual detection in the city, it will cause serious obstacles to the further processing and analysis of the results. Meanwhile, our method can achieve 96.3% accuracy with a *p*-value of 0.977, an R-value of 0.932, and a speed of 40.16 FPS. All of them are improved over the original YOLOv7 network, while the accuracy of IA-YOLO is 93.8%. Therefore, our method can be better applied in real-time UAV detection tasks. Our approach provides strong technical support and assurance for UAM.

In order to further prove our model for foggy UAV detection in an urban low-altitude complex background, a comparative experiment was conducted to analyze our model with the advanced model of the YOLO series. As shown in Figure 12a, the P-R curve comparison graph is shown, which reflects the relationship between the precision P and the recall R. The area enclosed by the recall and the precision represents detection accuracy. It can be seen that our method has the highest accuracy and the best performance of the model.

Figure 12b shows the F1-score curves of different models. The horizontal coordinate is the confidence level and the vertical coordinate is the F1, the closer to 1 means better performance. The F1-score of our method is 0.95. The top of the curve is closest to 1, indicating that our model can achieve a more balanced performance in the UAV detection task. Figure 13 shows the detection results of our model.

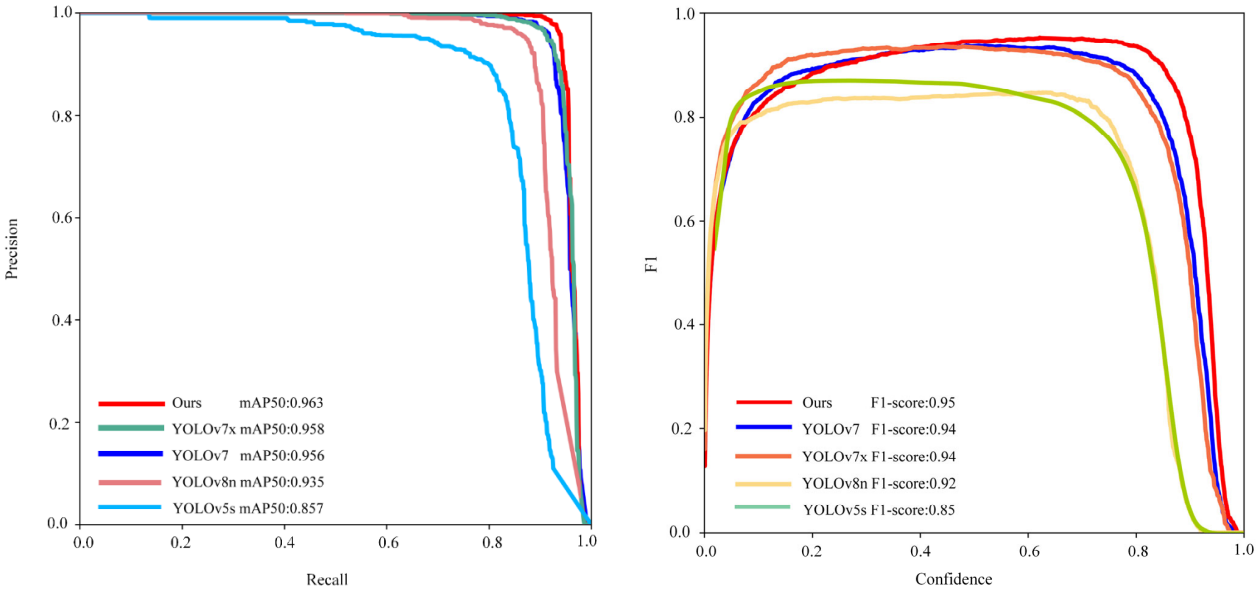


Figure 12. Comparison of P-R curves and F1-score curves of different models under visible conditions: (**a**) P-R curves; (**b**) F1-score curves.



Figure 13. Detection results on the UAV–visible dataset.

5.4. Experimental Analysis of Infrared Images

The results are analyzed in Table 5 for training and testing the UAV–infrared dataset. The YOLOv7 has an accuracy of 89.1% and a real-time performance of 32.79 FPS. In the results of the experiments comparing the improvement of MPConv in the backbone network, the results are analyzed according to the perspective of the accuracy of the detection results. The MPConv-GAM has the highest accuracy of 90.1%, followed by the MPConv-SA and MPConv-SE, which are 90.0% and 89.8%, respectively, all of which are improved compared to the original model. Analyzed from the perspective of real-time detection, MPConv-SE and MPConv-SA can meet the real-time detection requirement of 25 FPS. MPConv-GAM is 18.35 FPS, which cannot meet real-time detection requirements. Therefore, it can be seen that the improved MPConv method can not only enhance detection accuracy but also meets real-time requirements.

Table 5. Comparison of the results of different structural ablation experiments on the UAV–infrared dataset.

Model	MPConv-SE	MPConv-GAM [35]	MPConv-SA [34]	SPD-Conv	Precision (%)	Recall (%)	mAP50 (%)	Speed (FPS)
YOLOv7					90.0	81.9	89.1	32.79
	√				91.8	78.5	89.8	31.35
		√			88.6	83.8	90.1	18.35
			√		92.4	82.5	90.0	29.50
				√	96.4	88.1	93.8	30.03
			√	√	95.8	85.7	92.4	14.08
		√		√	94.4	86.2	92.1	26.89
	√			√	**97.1**	**89.7**	**94.7**	**28.57**

It can be observed that adding the SPD-Conv module to the small object head of YOLOv7 is able to achieve an accuracy of 92.1% with a real-time performance of 26.89FPS. It can be seen that the SPD-Conv is able to better detect UAV objects by objecting the low-resolution infrared image characteristics. Based on this model, an improved method is fused to the MPConv module. We can see from the table that adding the SPD-Conv module on top of the MPConv-SA module results in an accuracy of 92.4% and a real-time performance of 14.08 FPS. When we add the SPD-Conv module on the basis of the MPConv-SE module, the accuracy is 94.7% and the real-time performance is 28.57 FPS.

Figure 14 shows that our method achieves better detection in complex background conditions such as trees, tall buildings, and sky. We know from the dataset that most of the UAVs are small UAVs, and it is able to accurately detect UAVs under complex background conditions. This proves that our method can be applied in real detection tasks under the premise that accuracy and real-time performance are guaranteed.

We conducted multiple experiments on the same dataset to avoid randomness and ensure the reliability of our results. As shown in Table 6, the performance results of different modeling algorithms on the UAV–infrared test set are displayed. The comparison is made in terms of three metrics: P, R, and mAP50. Most of the existing more popular network architectures have poor performance in object detection and recognition and cannot meet the accuracy requirements for detection. In contrast, our method detects infrared images with an input size of 640 $\times$ 640 at 28.57 FPS with 94.1% accuracy. Compared with other methods, our method has better overall performance.



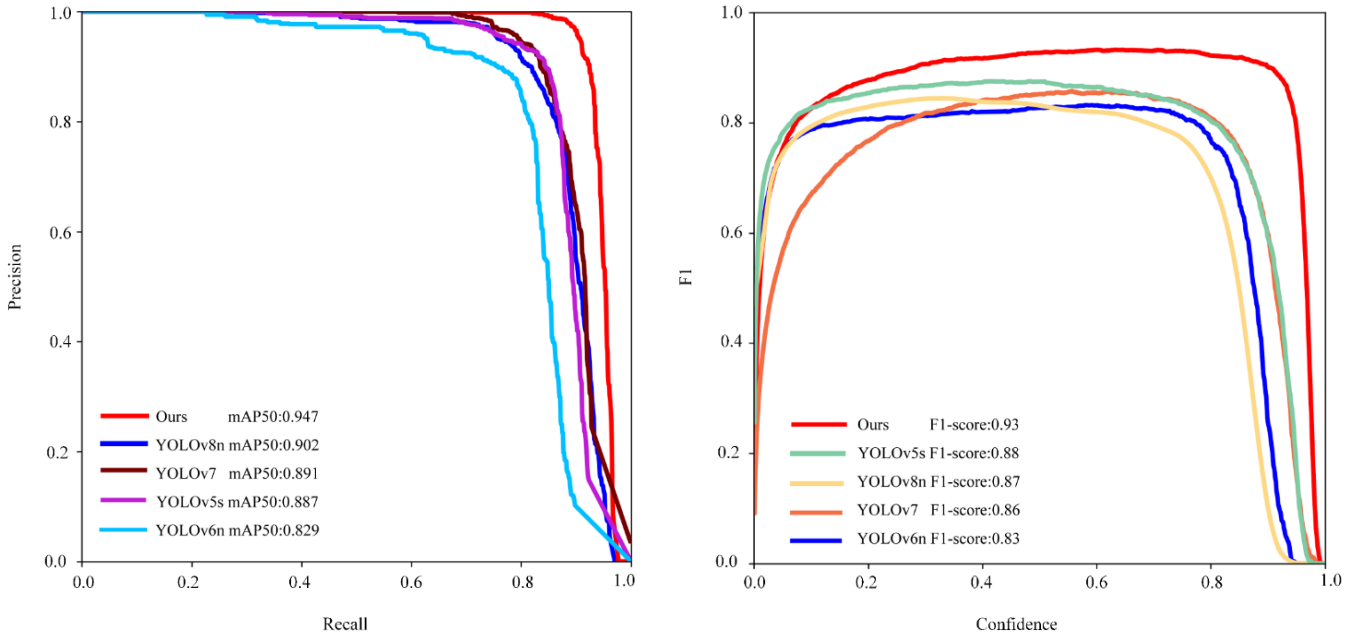
Figure 14. Detection results on the UAV–infrared dataset.

Table 6. Comparison of detection results of different models in infrared.

Model	Precision (%)	Recall (%)	mAP50 (%)
Faster-RCNN [36]	33.0	30.0	30.3
SSD [37]	60.5	68.8	64.0
YOLOv5s [8]	91.6	83.9	88.7
YOLOv6 [9]	80.2	84.4	82.9
YOLOv8n [11]	92.8	82.4	90.2
YOLOv7 [10]	90.0	81.9	89.1
Ours (visible)	**97.1**	**89.7**	**94.7**

Figure 15a shows the comparison of P-R curves, compared to other methods our method has the highest accuracy and the best performance of the model. Figure 15b shows the comparison of F1-score curves of different models. The F1-score of our method is 0.93. Compared with other models, the top of its curve is closest to 1, indicating that our model can achieve a more balanced performance in UAV detection tasks.

Our all-time detection model achieves UAV image detection during the day and night. In the day, the foggy weather detection problem is addressed by a defogging detection network. At night, the MPConv-SE and SPD-Conv structures are employed to detect low-resolution infrared UAV images. In summary, our algorithm is capable of real-time detection of Unmanned Aerial Vehicles in the day and night.



(**a**) (**b**)

Figure 15. Comparison of P-R curves and F1-score curves of different models under infrared conditions: (**a**) P-R curves; (**b**) F1-score curves.

6. Conclusions

An all-time real-time detection model is proposed by our team for the UAV detection tasks in urban low-altitude complex environments. In visible conditions, the model fuses the defogging structure to solve the problem on foggy days. In infrared conditions, the model fuses SE attention and converts space feature maps into depth feature maps. This structure optimizes the step convolution and pooling layers to solve the small UAV object detection problem. It can achieve great results for real-time detection tasks in complex urban environments. It is shown through extensive ablation experiments that our method performs well in all-time UAV detection. We also construct two UAV datasets in the complex context of urban low altitudes based on an adaptive approach. The all-time detection algorithm is trained separately on these two datasets, which can achieve 96.3% and 94.7% mAP50 on the UAV–visible and UAV–infrared datasets and perform real-time object detection with inference speeds of 40.16 FPS and 28.57 FPS, respectively. Our method can perform real-time detection for urban low-altitude complex background UAV objects night or day. In future research, we will delve deeper into UAV video surveillance methodologies. By considering the relationships between consecutive frames during the model detection process, we aim to continuously track the same object across frames. This will allow us to effectively leverage the spatiotemporal correlations between frames to enhance detection accuracy and efficiency. We hope that our work can benefit related scholars or academics to provide more technical support in the analysis of urban air traffic object detection applications.

Author Contributions: Conceptualization, Y.H. and J.Q.; methodology, Y.H. and J.Q.; software, Y.H. and H.W. and J.Y.; validation, Y.H., H.W. and J.Y.; writing—original draft preparation, Y.H. and J.Q.; writing—review and editing, Y.H., J.Q., H.W. and J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Tianjin Key Lab of Advanced Signal Processing and was supported by the Aeronautical Science Foundation of China under Grant 2022Z071067002.

Data Availability Statement: Data are contained within the article.

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

References

1. National Air Traffic Control Committee, Interim Regulations on the Administration of UAV, 2023. Available online: https://www.gov.cn/zhengce/zhengceku/202306/content_6888800.htm (accessed on 31 May 2024).
2. Civil Aviation Administration of China, Civilian Unmanned Aerial Development Roadmap V1.0 (Draft for Comments), 2022. Available online: https://www.caac.gov.cn/big5/www.caac.gov.cn/PHONE/HDJL/YJZJ/202311/P020231108392097578857.pdf (accessed on 2 November 2023).
3. Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; Torralba, A. Semantic Understanding of Scenes Through the Ade20k Dataset. *Int. J. Comput. Vis.* **2019**, *127*, 302–321. [CrossRef]
4. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
5. Huang, G.; Liu, Z.; Kilian, Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
6. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobile Nets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
7. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shuffle Net: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6848–6856.
8. Glenn, J.; Alex, S.; Jirka, B. YOLOv5s: v5.0-YOLOv5s-P6 1280 Models. AWS, Supervise.ly, and You Tube Integrations. 2022. Available online: https://github.com/ultralytics/ (accessed on 10 May 2023).
9. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. *arXiv* **2022**, arXiv:2209.02976.
10. Chien, W.; Alexey, B.; Hong, M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-time Object Detectors. *arXiv* **2022**, arXiv:2207.02696.
11. Dillon, R.; Jordan, K.; Jacqueline, H.; Ahmad, D. Real-Time Flying Object Detection with YOLOv8. *arXiv* **2023**, arXiv:2305.09972.
12. Wang, R.; Yan, J.; Yang, X. Learning Combinatorial Embedding Networks for Deep Graph Matching. In Proceedings of the 2019 IEEE International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2261–2269.
13. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 2017 Conference and Workshop on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017. [CrossRef]
14. Rao, Y.; Zhao, W.; Tang, Y.; Zhou, J.; Lim, S.N.; Lu, J. HorNet: Efficient High-Order Spatial Interactions with Recursive Gated Convolutions. In Proceedings of the 2022 European Conference on Computer Vision (ECCV), Tel-Aviv, Israel, 23–27 October 2022. [CrossRef]
15. Shaker, A.; Maaz, M.; Rasheed, H.; Khan, S.; Yang, M.H.; Khan, F.S. SwiftFormer: Efficient Additive Attention for Transformer-based Real-time Mobile Vision Applications. In Proceedings of the 2023 IEEE International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023. [CrossRef]
16. Brandon, Y.; Gabriel, B.; Quoc, V.; Jiquan, N. CondConv: Conditionally Parameterized Convolutions for Efficient Inference. *arXiv* **2019**, arXiv:1904.04971.
17. Yu, X.; Gong, Y.; Jiang, N.; Ye, Q.; Han, Z. Scale match for tiny person detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Los Alamitos, CA, USA, 2–5 March 2020; pp. 1257–1265.
18. Ye, T.; Zhang, J.; Li, Y.; Zhang, X.; Zhao, Z.; Li, Z. CT-Net: An Efficient Network for Low-Altitude Object Detection Based on Convolution and Transformer. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 2507412. [CrossRef]
19. Zeng, N.; Wu, P.; Wang, Z.; Li, H.; Liu, W.; Liu, X. A small-sized Object Detection Oriented Multi-Scale Feature Fusion Approach with Application to Defect Detection. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 3507014. [CrossRef]
20. Liu, S.; Wu, R.; Qu, J.; Li, Y. HDA-Net: Hybrid Convolutional Neural Networks for Small Objects Recognization at Airports. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 2521314. [CrossRef]
21. Minaeian, S.; Liu, J.; Son, Y.J. Vision-Based Object Detection and Localization via a Team of Cooperative UAV and UGVs. *IEEE Trans. Syst. Man Cybern. Syst.* **2016**, *46*, 1005–1016. [CrossRef]
22. Zhang, H.; Shao, F.; He, X.; Zhang, Z.; Cai, Y.; Bi, S. Research on Object Detection and Recognition Method for UAV Aerial Images Based on Improved YOLOv5. *Drones* **2023**, *7*, 402. [CrossRef]
23. Zhang, Z. Drone-YOLO: An Efficient Neural Network Method for Target Detection in Drone Images. *Drones* **2023**, *7*, 526. [CrossRef]
24. Huai, H.; Chen, Y.; Jia, Z.; Lai, F. Multi-object Detection and Tracking Algorithm for Urban Complex Environments of Intelligent Vehicles Incorporating Historical Trajectories. *J. Xi'an Jiaotong Univ.* **2018**, *52*, 132–140.
25. Liu, W.; Ren, G.; Yu, R.; Guo, S.; Zhu, J.; Zhang, L. Image-Adaptive YOLO for Object Detection in Adverse Weather Conditions. In Proceedings of the 2022 American Association for Artificial Intelligence(AAAI), Vancouver, BC, Canada, 2–9 February 2021. [CrossRef]
26. Kalwar, S.; Patel, D.; Aanegola, A.; Konda, K.R.; Garg, S.; Krishna, K.M. GDIP: Gated Differentiable Image Processing for Object Detection in Adverse Conditions. *arXiv* **2022**, arXiv:2209.14922.

27. Mazin, H.; Hayder, R. Multiscale Domain Adaptive YOLO for Cross-Domain Object Detection. *arXiv* **2021**, arXiv:2106.01483.
28. Liu, X.; Ma, Y.; Shi, Z.; Chen, J. GridDehazeNet: Attention-Based Multi-Scale Network for Image Dehazing. In Proceedings of the 2019 IEEE International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27–28 October 2019. [CrossRef]
29. Raja, S.; Tie, L. No More Strided Convolutions or Pooling: A New CNN Building Block for Low-Resolution Images and Small Objects. *arXiv* **2022**, arXiv:2208.03641.
30. Zhao, X.; Zhang, W.; Zhang, H.; Zheng, C.; Ma, J.; Zhang, Z. ITD-YOLOv8: An Infrared Target Detection Model Based on YOLOv8 for Unmanned Aerial Vehicles. *Drones* **2024**, *8*, 161. [CrossRef]
31. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal.* **2019**, *42*, 2779.
32. Liu, S.; Wu, R.; Qu, J.; Li, Y. HPN-SOE: Infrared Small Target Detection and Identification Algorithm Based on Heterogeneous Parallel Networks with Similarity Object Enhancement. *IEEE Sens. J.* **2023**, *23*, 13797–13809. [CrossRef]
33. Wu, H. *Research on Video-Based All-Time Vehicle Detection Method*; Hunan University of Technology: Zhuzhou, China, 2019.
34. Zhang, L.; Yang, Y. SA-Net: Shuffle Attention for Deep Convolutional Neural Networks. In Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 2235–2239.
35. Liu, Y.; Shao, Z.; Hoffmann, N. Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions. *arXiv* **2021**, arXiv:2112.05561.
36. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
37. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 21–37.

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