remote sensing

MDPI

Article

Visual Detail Augmented Mapping for Small Aerial Target Detection

Jing Li [1,*], Yanran Dai [1], Congcong Li [1], Junqi Shu [1], Dongdong Li [2], Tao Yang [2,*] and Zhaoyang Lu [1]

[1] School of Telecommunications Engineering, Xidian University, Xi'an 710071, China; yrdai@stu.xidian.edu.cn (Y.D.); ccli@stu.xidian.edu.cn (C.L.); jqshu@stu.xidian.edu.cn (J.S.); zhylu@xidian.edu.cn (Z.L.)

[2] SAIIP School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China; 1051196347@nwpu.edu.cn

* Correspondence: jinglixd@mail.xidian.edu.cn (J.L.); tyang@nwpu.edu.cn (T.Y.); Tel.: +86-139-9132-0168 (J.L.); +86-150-0291-9079 (T.Y.)

Received: 23 October 2018 ; Accepted: 19 December 2018 ; Published: 21 December 2018

Abstract: Moving target detection plays a primary and pivotal role in avionics visual analysis, which aims to completely and accurately detect moving objects from complex backgrounds. However, due to the relatively small sizes of targets in aerial video, many deep networks that achieve success in normal size object detection are usually accompanied by a high rate of false alarms and missed detections. To address this problem, we propose a novel visual detail augmented mapping approach for small aerial target detection. Concretely, we first present a multi-cue foreground segmentation algorithm including motion and grayscale information to extract potential regions. Then, based on the visual detail augmented mapping approach, the regions that might contain moving targets are magnified to multi-resolution to obtain detailed target information and rearranged into new foreground space for visual enhancement. Thus, original small targets are mapped to a more efficient foreground augmented map which is favorable for accurate detection. Finally, driven by the success of deep detection network, small moving targets can be well detected from aerial video. Experiments extensively demonstrate that the proposed method achieves success in small aerial target detection without changing the structure of the deep network. In addition, compared with the-state-of-art object detection algorithms, it performs favorably with high efficiency and robustness.

Keywords: small target detection; aerial video; visual detail augmented mapping

1. Introduction

Aerial target detection, as the key and foundation of avionics data understanding has a crucial impact on the whole system's performance, especially regarding the detection accuracy of small objects. The size of small objects occupies less than 1.0% of the total pixels. Compared with normal visual data, aerial data have unique characteristics in many respects. Their field of view is large in many cases and contains more visual content. Although it provides more comprehensive scene information for global analysis, the objects of interest usually account for less and do not have enough detail for detection. This leads to the failure of most state-of-the-art deep detection models. Some fail detection examples of the You Only Look Once Version 2 (YOLO v2) deep network are shown in Figure 1 (left). Therefore, effectively detecting small targets is one of the critical problems for aerial object detection systems. This challenging research topic has aroused wide interest in scientific and industrial circles due to its extensive application, including in large field monitoring systems [1–3], space-based early warning systems [4], territorial visual navigation [5,6], and so on.

However, small target detection from aerial video is much more difficult than normal object detection. The main reasons for this are described by the following points: (1) aerial video has a broad looking area that contains multiple background interferences for object detection. In aerial images, a great mass of the whole view is unrelated background, such as grassland, trees, buildings, etc. Meanwhile, the small target size is not sufficient for accurate object representation. (2) Aerial targets vary greatly in size due to their different flight heights and camera angles. Besides that, small aerial targets have poor-quality appearances and structures in most cases. That leads to difficulty extracting object features well from limited data. (3) Moving targets in aerial video can be easily confused with noise, due to their small sizes. That leads to a proportionate increase in the detection error rate. The factors above bring certain difficulties to small target detection from aerial data.

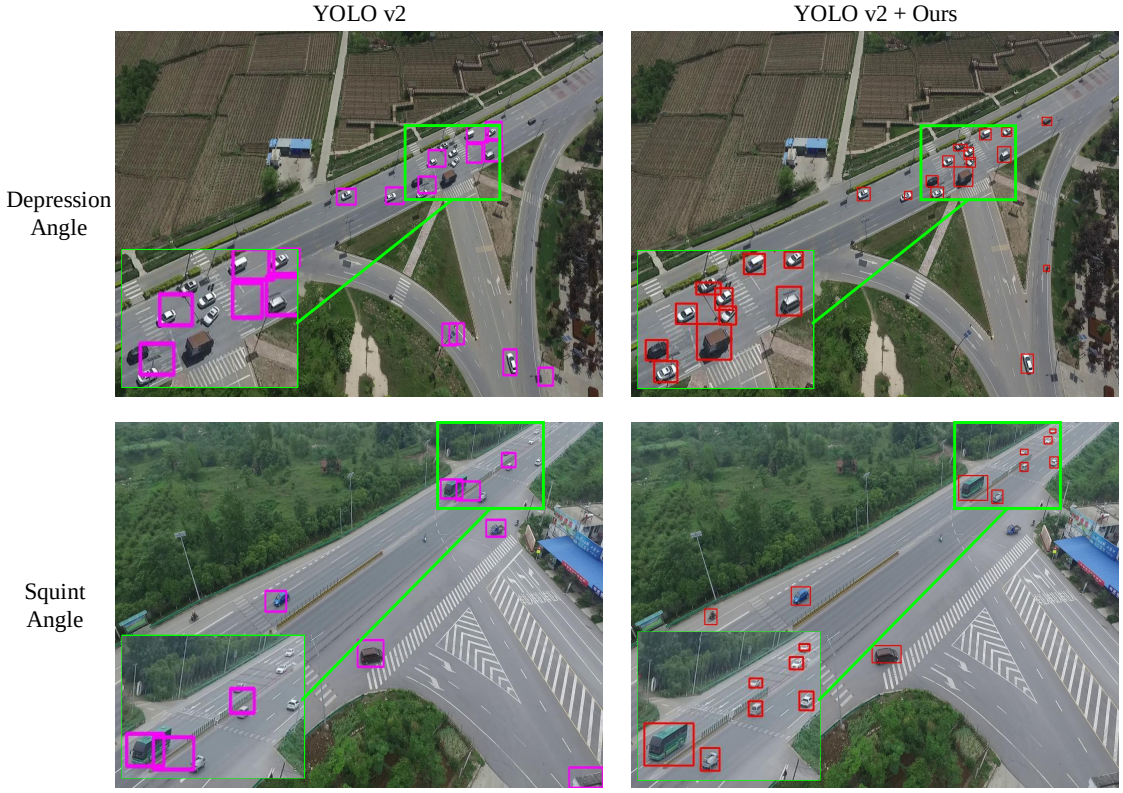

Figure 1. Some small object detection results with You Only Look Once Version 2 (YOLO v2) deep network and the proposed method. (**left**) The detection results with the YOLO v2 deep network; (**right**) The detection results with YOLO v2 and our visual detail augmented mapping approach. This figure shows that the proposed method achieves better performance in terms of both precision and recall.

To date, some effort has been devoted to addressing the problem of small object detection from aerial video over the past decade [7–10]. One widely applied strategy is to directly enlarge images to different scales. This kind of approach achieves more detailed information of small targets by magnification. For example, Chen et al. [11] presented an approach where the input is magnified to enhance the resolution of small objects. On the basis of this research idea, Cao et al. [12] fused feature modules to additional contextual information to achieve better detection performance. By generating multiple feature maps with different resolutions, they were able to naturally handle objects of various sizes including those of small sizes. However, with an increase in image size, these approaches are accompanied by large computation costs. That means that they cannot meet the real-time requirements for practical applications. Other proposed approaches are based on the deep neural network in which each small target characteristic is represented by multi-scale feature layers [3,13,14].

In 2015, Li et al. [13] presented a novel deep network based on the Viola–Jones framework [15], and their approach achieved great face detection performance. Region proposal-based detection networks have been extensively applied in the target detection field, including Region-Convolutional Neural Networks (R-CNN) [16], Fast-Region Convolutional Neural Networks (Fast-RCNN) [17], and Faster-Region Convolutional Neural Networks (Faster-RCNN) [18]. Through analyzing these works, scientists put forward some small target detection algorithms [19–21]. Moreover, Cai et al. [14] proposed a Multi-scale Deep Convolutional Neural Network (MS-CNN) which predicts objects at different layers of the feature hierarchy. However, the high performance of these approaches deeply depends on training data. It cannot grant the discriminability of small target features because rich representations are difficult to learn due to their poor-quality appearance and structure. All in all, the approaches of magnifying images and using the deep learning network algorithms have their own drawbacks on small object detection from aerial video.

To address this issue, we firstly analyze the reason for the failure of small target detection. This is mainly due to two aspects of the characteristics of small targets. On the one hand, small targets are difficult to describe as they have less counterpart pixels in aerial video. That is the ultimate cause of error detection. On the other hand, small aerial targets are easily confused with noise. This is because the background of aerial video is dynamic in most cases and it leads to a high false alarm rate. Through the discussion above, it was found that the key to solving the problem is to represent a small target precisely and concretely. As is universally acknowledged, better object representation depends on the support of sufficient pixel data. In a sense, the first idea that comes into our minds is to enlarge images on multiple scales to obtain detailed information about potential small objects.

As Figure 2 shows, we experimentally evaluate detection performance with multiple scale processing. To ensure fairness and objectivity, all experiments are conducted under the same deep detection model. After multi-resolution, the detection performance on small targets is greatly improved as compared with direct detection. However, at the same time, the computation cost is increased greatly and cannot meet real-time demands. For proper detection systems, computing time and detection accuracy are both significant. In other words, the key to implementing a robust small target detection system is obtaining detailed target information within a short processing time. Therefore, considering the detection accuracy and efficiency, we aim to establish a mapping relationship. Based on this, original small target regions will be mapped to new foreground space that contains more valuable object information for better representation. Naturally, the detection performance is improved with the premise of a few additional calculations and no change to the deep network.

In this paper, we systematically investigate the above-mentioned fundamental idea and propose a novel visual detail augmented mapping algorithm for small aerial target detection. To be specific, the proposed approach can be divided into three parts: multi-cue foreground segmentation, visual detail augmented mapping, and small object detection with a deep network. The first part synthetically analyzes the motion information and graychange information of the moving target, and extracts the potential regions from the input aerial video. The locations of these regions should be visually enhanced. Then, we put forward a visual detail augmented mapping approach. On the basis of this mapping, we magnify the potential target regions to multiple resolution to provide detailed information and rearrange them into new foreground space for visual enhancement. Thus, the original small targets are mapped to a more efficient foreground augmented map which is favorable for accurate detection. With the proposed visual detail augmented mapping, not only can the interference of unrelated area be further filtered out, but also, more detailed target information can be obtained by a subsequent detection network. Finally, driven by the success of the deep learning network YOLO v2, which is pre-trained by normal-sized objects, a small target detection system is implemented with respect to accuracy and speed. Figure 1 (right) shows some examples of detection results.

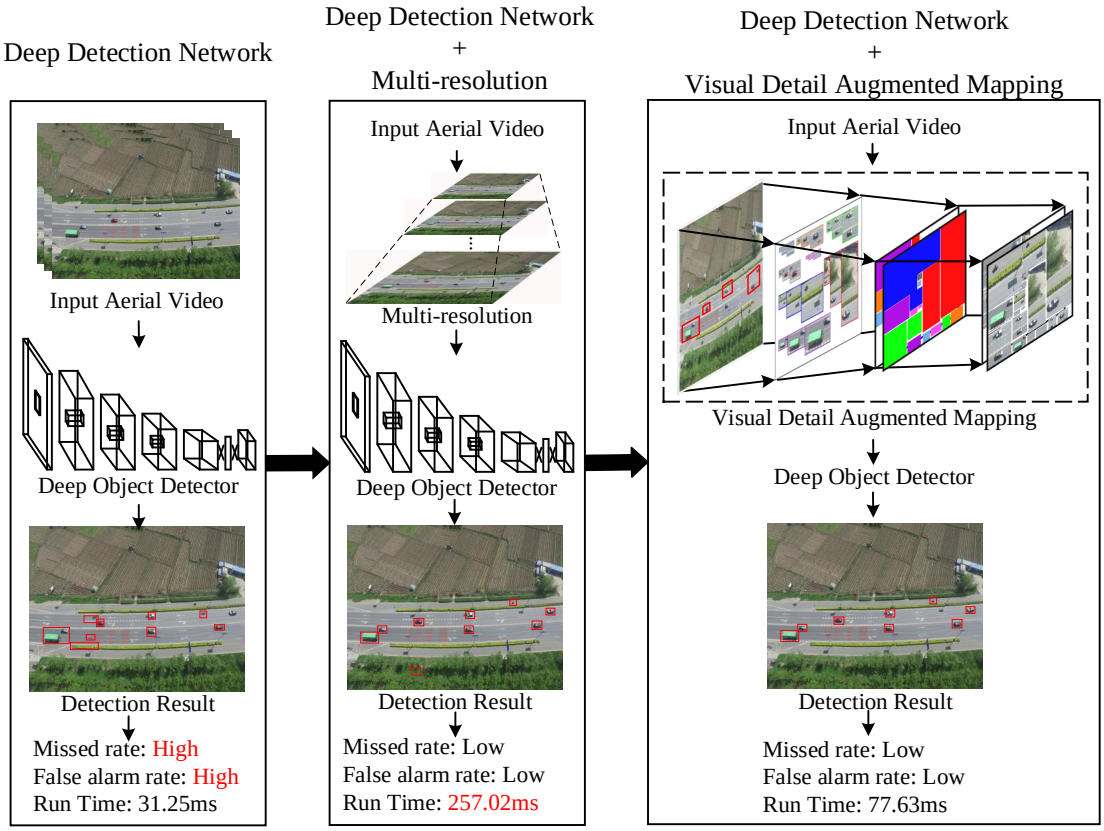

Figure 2. The research route of the proposed method. This paper investigates the problem of small aerial object detection. The general solution is based on the deep detection model, but direct detection causes high false alarm and miss rates. Another approach is to add multi-resolution processing before deep detection. However, this improves the detection performance and reduces the efficiency at the same time. In order to balance effectiveness and efficiency, this paper proposes a novel visual detail augmented mapping approach that maps small target regions to a new foreground space to allow better detection.

The main contributions of our work can be summarized as follows:

- We propose a novel visual detail augmented mapping approach that provides a wealth of specific information about the small target of interest. Aerial targets usually have few counterpart pixels and are hard to describe. Through mapping, these potential regions are mapped into a new foreground space with more abundant small target information. It has been proved that our approach based on visual detail augmented mapping offers a more valuable foreground augmented map for subsequent detection network, and it is greatly beneficial as it achieves small target detection without changing the configuration or framework of the deep network.
- We present a multi-cue foreground segmentation method to extract interesting regions. These potential regions might contain target information that needs to be enhanced. Thus, through the visual detail augmented mapping approach, input aerial video is selectively mapped to a new foreground space. The small target detection system is well implemented with the deep detection network in the new foreground space. Experimental results indicate that the proposed method greatly improves the detection performance for small targets with a small increase in the computation load.
- Based on visual detail augmented mapping, we propose a small aerial target detection algorithm. Concretely, our training database is established by combining a self-built database with the public database UA-DETRAC. To better evaluate the performance of the proposed method, we also

build a test database that captures different aerial scenes, including an intersection, a T-junction, a road beside a parking lot, and a twin multi-lane urban road. Experiments demonstrate that the proposed method shows encouraging results, and it provides an improvement in performance of greater than 15% compared with direct deep network detection.

The paper is organized as follows. Section 2 presents a general overview of the proposed algorithm and describes each detection part in detail. The experiment results are shown in Section 3. Section 4 discusses the performance of the proposed approach with other methods. Finally, we conclude this paper in Section 5.

2. Visual Detail Augmented Mapping for Small Target Aerial Detection

In this section, the proposed small target detection algorithm is introduced in detail. First of all, the potential target is extracted by multi-cue foreground segmentation algorithm, and interesting areas are approximately divided into groups. This establishes the foundation for follow-up detection. After that, we propose a novel visual detail augmented mapping method to map these potential regions into a new foreground space, which is the fundamental technique of our work. Finally, using the pre-trained deep detection network, a system for small target detection from aerial video is implemented. Figure 3 shows an overview of the proposed method.

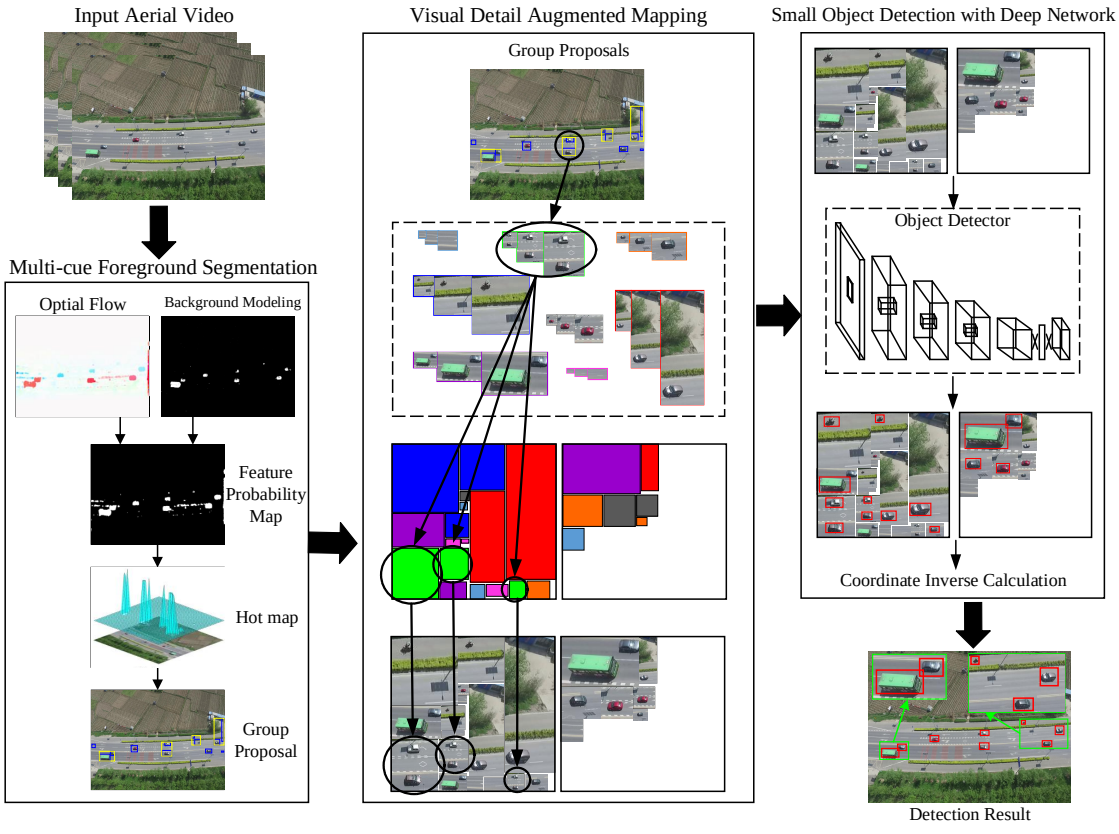

Figure 3. An illustration of the proposed method. The small target detection method contains three parts: multi-cue foreground segmentation, visual detail augmented mapping, and small object detection with the deep network. After the three modules, small targets can be detected accurately and quickly from the input aerial video.

2.1. Multi-Cue Foreground Segmentation

The main problem when trying to detect small moving objects from aerial video is to separate the changes in the image caused by objects from those caused by the dynamic background. Therefore, at the beginning, we present a multi-cue foreground segmentation method to extract the potential target region that needs visual detail augmented mapping. As Figure 3 shows, the proposed method comprehensively analyzes the optical flow and background changes that might be caused by moving objects. Through the process above, its feature probability map is obtained by combining these two pieces of information. Then, in order to reduce the noise effect, the probability map is smoothed and its corresponding hot map is generated. The hot map symbolically depicts the possible position of the target. Then, single potential targets are clustered into group proposals and follow-up detection algorithm is processed in group.

As for multi-cue object information, optical flow and background modeling are adopted to jointly obtain the target probability map. The choice of these two algorithms is mainly based on two points: (1) Optical flow is an important technology for motion estimation, as it represents the relative motion information for each pixel. Optical flow can help us to pinpoint small targets more accurately. (2) The background modeling method describes grayscale changes between the current image and background by modeling the image background. This method is sensitive to small and weak changes and therefore is suitable for small target detection. However, both approaches have their merits and shortcomings. Optical flow is not robust on small targets and might result in missed detections. Background modeling can be easily disturbed by noise which may be mistaken for small targets, leading to a high false alarm rate. Though the two methods have their own limitations, there is a complementary role between them. The background modeling method will produce a miscarriage of justice when there is only a little gray contrast between the target and background, while optical flow can maintain good performance because its similar background also has a certain degree of motion. However, the optical flow method is insensitive to weak motion, which is advantageous for the background modeling method. In other words, optical flow and background modeling are different ways to describe the moving object. This means that the optical flow method combined with background modeling will have dual advantages and gain better effects for small target detection from aerial video. Based on the above analysis, the proposed method utilizes the advantages of the two methods to get the potential object region.

In the process of implementation, Farneback [22] is first employed to calculate the optical flow information on the basis of considering detection speed and accuracy. The Farneback algorithm was proposed by Gunnar Farneback in 2003. It is a global dense optical flow algorithm based on two-frame motion estimation. Suppose that the two adjacent frames to be detected are denoted as I_{t-1} and I_t, respectively. Firstly, the coefficient vector of each pixel is calculated by the polynomial expansion transform. Taking pixel point $\mathbf{x}$ on I_{t-1} as an example, the approximate position $\widetilde{\mathbf{x}}$ of this pixel on the next frame image I_t is

$$\widetilde{\mathbf{x}} = \mathbf{x} + \widetilde{d}(\mathbf{x}). \tag{1}$$

If the parameter of I_{t-1} is $\mathbf{A}_{t-1}(\mathbf{x})$, $\mathbf{b}_{t-1}(\mathbf{x})$ and I_t is $\mathbf{A}_t(\mathbf{x})$, $\mathbf{b}_t(\mathbf{x})$, the intermediate variable $\mathbf{A}(\mathbf{x})$, $\Delta\mathbf{b}(\mathbf{x})$ can be formulated as

$$\begin{aligned} \mathbf{A}(\mathbf{x}) &= \frac{\mathbf{A}_{t-1}(\mathbf{x}) + \mathbf{A}_t(\widetilde{\mathbf{x}})}{2}, \\ \Delta\mathbf{b}(\mathbf{x}) &= -\frac{1}{2}(\mathbf{b}_t(\widetilde{\mathbf{x}}) - \mathbf{b}_{t-1}(\mathbf{x})) + \mathbf{A}(\mathbf{x})\widetilde{\mathbf{d}}(\mathbf{x}). \end{aligned} \tag{2}$$

Thus, we can get the coefficient vectors $\mathbf{G}(\mathbf{x})$ and $\mathbf{h}(\mathbf{x})$ by further calculations. Suppose that $\mathbf{S}(\mathbf{x})$ is the scale matrix. The optical flow $\mathbf{d}(\mathbf{x})$ is directly solved by

$$\begin{aligned} \mathbf{G}(\mathbf{x}) &= \mathbf{S}(\mathbf{x})^T \mathbf{A}(\mathbf{x})^T \mathbf{S}(\mathbf{x}) \mathbf{A}(\mathbf{x}), \\ \mathbf{h}(\mathbf{x}) &= \mathbf{S}(\mathbf{x})^T \mathbf{A}(\mathbf{x})^T \Delta \mathbf{b}(\mathbf{x}), \\ \mathbf{d}(\mathbf{x}) &= \mathbf{G}(\mathbf{x})^{-1} \mathbf{h}(\mathbf{x}). \end{aligned} \tag{3}$$

The Farneback method obtains the object's moving information by analyzing its optical flow transformation, as shown in Figure 4a. Its high detection accuracy and rapid data processing speed fit the requirements of a real-time aerial object detection system.

As for the background modeling method, the Fast-MCD algorithm [23], which was proposed by Kwang et al. in 2013, was used. It segments the moving foreground by modeling the image background through the dual-mode Single Gaussian Model (SGM) with age. This method was chosen for two reasons: its robustness for testing moving backgrounds which frequently appear in aerial videos and its attention to small changes which tends to result in a lower missed detection rate on small targets.

Specifically, the procedure consists of the following steps. First of all, in order to remove background movement, Fast-MCD calculates a homograph matrix for the perspective transform from I_{t-1} to I_t. Then, the SGM model of the candidate background and the apparent background can be obtained. Assume that the input aerial image I_t is divided into several grids of the same size, the ith grid is denoted as $\mathbf{G}_i^{(t)}$ in this paper. Through analysing the mean, variance, age of the candidate model, and the apparent model, the background model is selectively updated. Finally, after obtaining the background model, the image foreground is detected. On the basis of the updated apparent model with the mean μ, variance σ, and threshold θ_b, the pixel p with gray scale T_p which satisfies Formula (4) is the foreground. Thus, the moving foreground which contains small targets is segmented from the complex background:

$$(T_p - \mu)^2 > \theta_b \sigma. \tag{4}$$

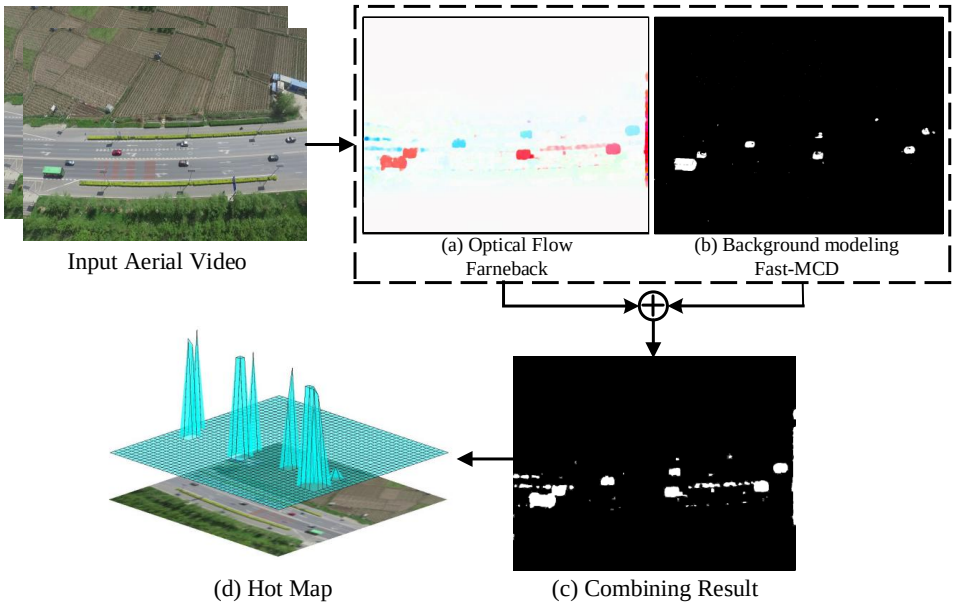

Figure 4. The probability result by combining optical flow and the background modeling method: (**a**) optical flow result; (**b**) background modeling result; (**c**) combined result; (**d**) hot map.

Figure 4a,b show the moving foreground segmentation results of the Farneback and Fast-MCD, respectively. These figures show that the two methods have their own advantages as well as drawbacks. On the one hand, the overall outline of Farneback is more complete, but it lacks the small target on the upper right corner, while Fast-MCD can accurately obtain each small gray scale change, which is the disadvantage of Farneback approach. On the other hand, there are some holes in the detection results of Fast-MCD. The production of imcomplete objects can be avoided by Farneback. Overall, the results show that the two algorithms can exert a complementary action in small object detection. From another perspective, the purpose of foreground segmentation is to enhance the target saliency relative to the background. Background modeling and optical flow start from two different views to describe an object's characteristics. The combination of the two methods actually improves object saliency and that helps us to segment a potential object region from the complex background. Thereafter, Farneback and Fast-MCD are combined to calculate the probability map. For the combined method, this paper takes the union of the two segmentation results to retain every possible position of the target, as shown in Figure 4c shown. The combination of the two methods can be regarded as its feature probability map. It not only combines the advantages of optical flow and Fast-MCD, but also overcomes the shortcomings of both to a certain extent.

With probability map, the following is the potential area identified. In this step, considering that noise will damage the detection performance, the proposed method first smooths it. Mean filtering is used as the image smoothing algorithm. The mean filter is one of the most commonly used linear filters whose output is a simple average of all neighborhood pixels. This algorithm reduces sharp changes in the probability map and some noise can be filtered out from the probability map. As for the fuzzy problem caused by the mean filter, this paper employs it to locate potential objects in the probability map rather than in the aerial image. Therefore, it will not cause the blurring of the image in follow-up links. The detailed calculation process of the hot map is as follows. Take the pixel with coordinates in the probability map g of (x_i, y_i) as an example, the filter window size is $S = m \times n$. The pixel value $f(x_i, y_i)$ of this point in the hot map is

$$f(x_i, y_i) = \frac{1}{mn} \sum_{(p,q) \epsilon S} g(p, q). \tag{5}$$

In this way, the hot map can be calculated from probability map with less noise impact. Figure 4d show the visualization results of the hot map. The larger the hot value is, the greater the probability that a moving object exists. On this basis, the foreground area is selected from the detected hot map with the minimal bounding box and a set of candidate bounding boxes is obtained. These positions are where small targets might appear.

However, overlaps exist between these bounding boxes. As for the small targets placed in the far source, they occupy too few pixels and these overlaps may split a target. Incomplete objects have a great impact on the detection performance. This problem lets us think of a human visual system. Imagine a scenario with trees and mountains, with a user admiring the view. To detect tree regions from this scene, the human visual system only needs to group the forest area roughly. Similarly, we just need to segment the potential small targets into a group for the aerial video to obtain the region of interest. Inspired by this idea, in this paper, we apply it to the moving foreground segmentation.

To be more specific, the initial bounding boxes are clustered based on the distances between them to get group proposals. This reduces the target fragmentation caused by overlap which brings in a few parts of the background. In the processs of implementation, there are two cases that need to be clustered. First, the proposed method merges these overlapped boxes into one bounding box. This reduces the situation where overlapped boxes split the real object and provides more precise information for the following detection. Second, bounding boxes that are close together are clustered. Rather than using single objects, the division into groups is more suitable for small target detection due to the small sizes of the objects. In addition, it is enough to confirm the approximate locations of small targets at the preprocessing stage, and then have accurate detection followed up by the algorithm.

In this way, the bounding boxes are synthesized into a group. These areas are the regions of interest in the aerial video. This has the benefit of improving the detection accuracy by avoiding the situation of small bounding boxes splitting the real target and reduces the complexity of subsequent computing with fewer areas of interest. Through the above processes, the group proposals that contain the small targets of interest are all segmented.

2.2. Visual Detail Augmented Mapping

So far, these regions of interest that might contain real small targets have been segmented from the input aerial video. The next step is to accurately detect whether these regions have moving objects or not. However, due to the small size of the aerial targets, direct detection will lead to a high false alarm rate and missed rate. To solve this problem, this section describes the corresponding analysis research and proposes a novel approach to solve this problem.

First of all, we explore the root cause of the small target detection failure. From the results obtained in the previous step, we make the following conclusion. Although the group proposal can basically frame all potential moving targets, it still contains some problems. The targets in an aerial image are so small that we cannot obtain enough information from them. The limited visual data makes it difficult for us to represent object well. This is the fundamental reason why the accuracy of small targets is low. To tackle this problem, the simplest solution is to enlarge the aerial image and get more detailed characteristics about the object representation. However, as the experimental results in Figure 2 show, the larger the image is, the more pixel level information it contains. This means that a number of calculations are required in many image processing links. In aerial video, unrelated background areas usually occupy a large proportion of the total area, and there is only a tiny area with the target of interest. That is to say, enhancement of only a few pixels is enough to achieve accurate target detection. According to this principle, in this paper, we present a visual detail augmented mapping algorithm approach which gives attention to both the detection speed and precision.

This mapping mainly includes two modules: one is multi-resolution mapping which selectively enlarges the potential region to different sizes to give detailed object information; the other is foreground augmented mapping which maps the original target to a more compact foreground space for visual augmentation. The framework is given in Figure 5.

In the first part, the proposed approach performs multi-resolution mapping on the group proposals obtained in the previous section. Let the input I_t contain N group proposals; its group proposal set is denoted as $\mathbf{R} = \{\mathbf{R}_1, \ldots \mathbf{R}_i, \ldots, \mathbf{R}_N\}$. $\mathbf{R}_i$ is the *ith* group proposal with coordinates $\{x_i, y_i, width_i, height_i\}$, where (x_i, y_i) are the coordinates of the upper left corner and $width_i \times height_i$ is the corresponding bounding box size. For one group proposal, the proposed method enlarges it into different scales. That can transform the small target in an aerial video into its normal size. For example, $\mathbf{R}_i$ covers an area of $width_i \times height_i$. We magnify it into three sizes, and the scale factors are 1.0, 2.0, 3.0. Thus, the size of $\mathbf{R}_i$ is increased to $1.0 \times width_i \times height_i$, $2.0 \times width_i \times height_i$, $3.0 \times width_i \times height_i$, respectively. For the amplification interpolation method, we utilize linear interpolation, which is a commonly used method. By enlarging all group proposals with linear interpolation to the three scales above, we obtain the amplified results of each region, as shown in Figure 5a. We can see from the figure that small targets in the input aerial image are converted into normal size through mapping at the three scales. More detailed target information is mapped out.

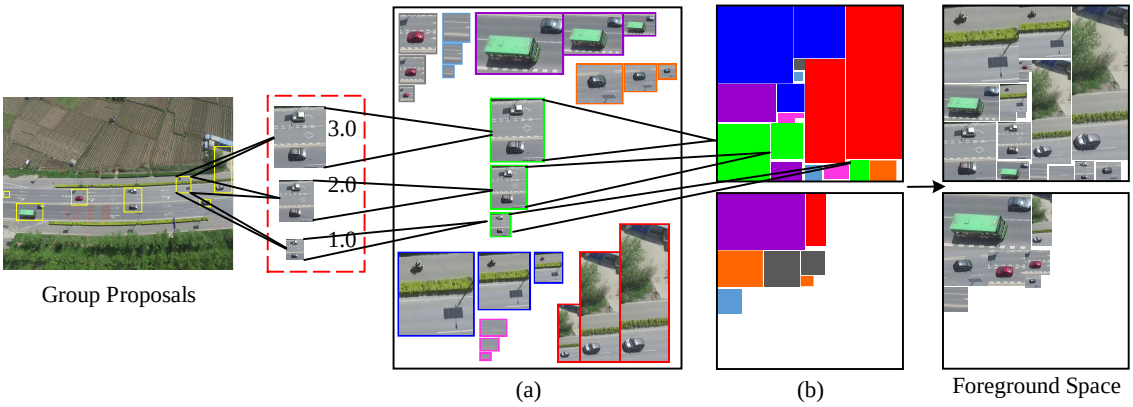

Figure 5. An overview of the visual detail augmented mapping method. This contains two modules: one is (**a**) multi-resolution mapping and the other is (**b**) foreground augmented mapping. Taking the group proposal involving two cars obtained by Section 2.1 as an example, through three scales of multi-resolution, the cars are mapped to normal size with more detailed object information. Then, the second part maps them into a new foreground space in which the original target area is enhanced.

Through the multi-resolution mapping processing, the specific characteristics of small targets are mined out which is advantageous to object representation and, in turn, strengthens the detection performance. For efficiency, we only map a small portion of the input image which takes little additional computation cost. In other words, our method improves the detection accuracy of small aerial targets and also gives attention to the detection speed.

After that, how do we integrate this detailed target information? The most simple and direct way is to send these enlarged regions to be detected one by one. This method requires as many detection times as the number of enlarged regions. This leads to a low detection efficiency. Another solution is to put all of them directly into a large enough foreground image, and then the large foreground image is sent once into the deep detection network. Though the efficiency of this method is high, its foreground images will be resized to a small scale and that will destroy the object's detailed information. Therefore, in this part, we present a novel foreground augmented mapping method which rearranges these regions into a efficient and compact space. Compared with the other two methods, foreground mapping provides a method to minimize the loss of efficiency while retaining the augmented visual effect. As Figure 5b shows, the proposed method maps the potential area with multi-resolution into a set of foreground augmented maps and ditches the irrelevant background part. Thus, it not only can reduce the influence of the background and avoid computing resources waste, but also further visually enhance the valuable detailed target information. To be specific, the proposed method first creates a set of empty images in the new foreground space. The size of foreground augmented map is designed to be dependent on the follow-up deep detection network. This ensures that the target regions are not reduced by the detector's internal steps, and there is no target information loss in this step. The number of these images is determined by potential region's size. Then, in order to make full use of each augmented map in the foreground space and to save the computation source, we pack as much of the region as possible into the limited image space. This issue can be regarded as the rectangular packing problem which is a combinatorial optimization problem. In this paper, we employ the rectangular packing algorithm to find the optimal solution for potential target region location. The packing result is shown in Figure 5b. As we can see, all regions of this instance are packed into two foreground augmented maps with maximum space utilization. The two foreground augmented maps are what we will send to the deep detection network. Through visual detail augmented mapping, the input aerial video is visually augmented in two ways: (1) Potential target regions are mapped into multiple resolution which enhances the detailed target features and provides a subsequent detection network with more abundant visual information. (2) The new foreground space filters out most of the

irrelevant background and only rearranges the potential target regions. In some sense, the foreground augmented maps are the visual augmentation of the original aerial data.

Now, after visual detail augmented mapping, the preprocessing work before detection is finished. The result of the procedure above is a set of foreground augmented maps which are mapped from the original aerial image and they only contain some valuable potential target regions which need further precise detection.

2.3. Small Object Detection with the Deep Network

Through the methods described the previous sections, a set of foreground augmented maps constructed by visual detail augmented mapping is obtained. All potential regions on different scales are well arranged in these foreground augmented maps. As Figure 6 shows, a large proportion of the foreground augmented map is the valuable object area and the unrelated background only occupies a small part. The reverse situation appears in the input aerial video. Based on this prospective image which is beneficial to detection, this section sends them to the deep detection network and outputs the final detection result. Concretely, this part is composed of two modules: preliminary small detection and coordinate back calculations.

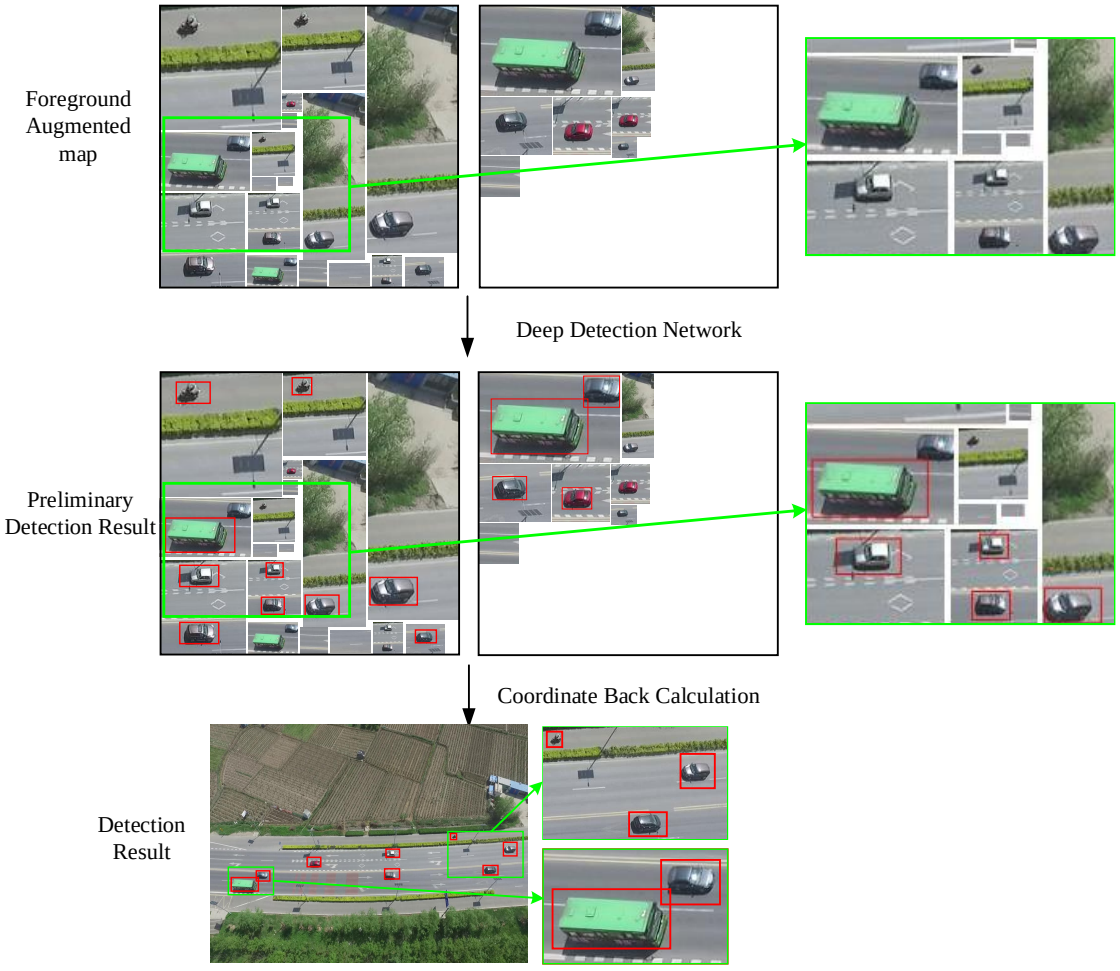

Figure 6. Small object detection with the deep network. First, on the basis of the YOLO v2 deep detector, targets are preliminarily detected. Second, we inversely calculate the coordinates in the foreground space into an input aerial image and get the final small object detection result.

For the first detection part, we employ You Only Look Once Version 2 (YOLO v2) [24] as the deep detection model. YOLO v2 is proposed by Joseph and Ali in 2017. It is a real-time object detection network which has achieved widespread success in most normal object detection tasks. YOLO v2 can detect over 9000 object categories, and it is robust to different tasks with fast processing speed and high detection accuracy. However, the main disadvantages of YOLO v2 is that its performance degrades badly when object size is small. Therefore, we apply visual detail augmentation mapping to overcome the limitations of a YOLO v2 deep detection network. The basic detection process of YOLO v2 is as follows: (1) The well rearranged foreground images are in turn sent to the pre-trained detection neural network. In general, the YOLO v2 detector is generated by training a large amount of data and the detector performance is heavily influenced by training data. However, with previous visual detail augmented mapping, our method does not need to change the detection network for a small target and it utilizes the pre-trained detector which is generated for normal size target detection. After multiple layers operation, the feature maps of each foreground augmented map with size 13×13 can be obtained. (2) On the basis of feature map, YOLO v2 predicts the bounding box and calculates their confidence in each category. (3) With the loss function, the trained network detection further screens these bounding boxes and the targets with high confidence can be considered as the real object. Thus, the object detection procedure is finished. The detection result which contains each object's location is shown in Figure 6. Though there is still a small proportion of background, almost all objects can be detected accurately in foreground augmented map at least one scale. It has proved the performance of the proposed method once again. In addition, the limitation of aerial data and long training time make it difficult to retrain an appropriative deep network for small object detection. In addition, our method with a pre-trained network for normal size object detection shows its great advantages in small target detection.

However, the positions of these detection results are coordinates in the mapped foreground space. Therefore, in the final step, we inverse calculate these coordinates. Specifically, the proposed method filters the bounding boxes which are the same objects on different scales at first. Only the bounding box with the highest overlapping rate is retained. Then, based on the mapping relation, we back calculate their coordinates on the initial aerial image and do the corresponding scaling. In this way, we finish all of the small target detection steps. The detection results are labeled on the input image as shown in Figure 6.

To summarize, based on the multi-cue foreground segmentation method and the visual detail augmented mapping algorithm, the follow-up deep detection network obtains more detailed and specific target information in the new mapped foreground space. With no need to change the framework of the detection network and with a small increase in computation, the proposed approach implements a small moving object detection system for the aerial video quickly and accurately. Compared with direct deep network detection, the detection performance of small targets is greatly improved.

3. Experiment Results and Analysis

Extensive experiments were conducted to fully evaluate the small moving object detection performance of the proposed method based on visual detail augmented mapping. In this section, we firstly introduce the aerial visual database that was used for training and testing. Then, we discuss the multi-cue foreground segmentation method. Finally, the qualitative and quantitative detection results of the proposed method on two different visual angles are presented.

The common configuration for all experiments is summarized here. The frame rate of all input aerial videos was 30 FPS and each frame size was 1280×960. The program was implemented in C++ and all results determined on an Intel (R) Core (Santa Clara, California, USA) (TM) i7-7700HQ (2.80 GHz CPU, 8 GB RAM).

3.1. Database

To test the advantages and generalization of the small moving object detection algorithm, experiments were performed on some challenging aerial scenes. This section respectively introduces the training database and test database for performance analysis.

Training database: The detection performance greatly depends on the training database. In order to improve the detection ability of small targets, the general resolution is to retrain a new deep network. However, the small object detection model involves many difficulties, mainly concerning two aspects: first, the number of public aerial databases is small and their scales are also not big enough; second, training a new network is not only time-consuming but also requires a high calculating capacity. Therefore, in this paper, we combined a public database and self-built aerial data to form our training database. As for the public database, we employed the UA-DETRAC [25] database which was shot on the road overpass in Beijing and Tianjin. We chose 11 representative scenarios with 10,963 images as our training database. In addition, we complemented the training database with aerial data captured by DJI M100 (Shenzhen, China), which contains 3175 images. The vehicle sizes in these images are normal. Therefore, we pretrained a YOLO v2 model with 14,138 images as our deep detection network.

Test database: Taking into account that most public databases are not generated for small aerial object detection, we established a new aerial test database which contained a number of complex traffic aerial scenes, including the twin multi-lane urban road, road beside parking lot, intersection, and T-junction. Figure 7 shows the data acquisition equipment and some examples from the test database. As we can see, all aerial data was captured by DJI M100 around different major transport arteries. Concretely, this database consisted of 18 images sequences formed by 79,742 frames. We divided them into five scenes with different shooting angles: depression and squint. For the depression angle, the average pixel proportion of each moving object was less than 1.0%. This is suitable for verifying the detection performance of small objects. The object size of the squint aerial video varied over a wide range from 150×150 to 10×10. This meets the demands of testing the robustness on different object sizes. Besides that, in order to conduct the quantitative and qualitative experimental analyses better, we manually annotated the ground truth to provide a moving object reference position. Concretely, we utilized LabelImg [26] as our label tool. This annotation tool is easy to use and can generate XML files directly. To make the statistical analysis meaningful, each scene included 2000 labeled images, and there were 7–8 moving vehicles in each frame, on average. As Figure 7 shows, we took the minimum bounding box on each moving vehicle as the ground truth, which is marked with a yellow box. On the basis of these reference coordinates, we evaluated the detection performance of the proposed method.

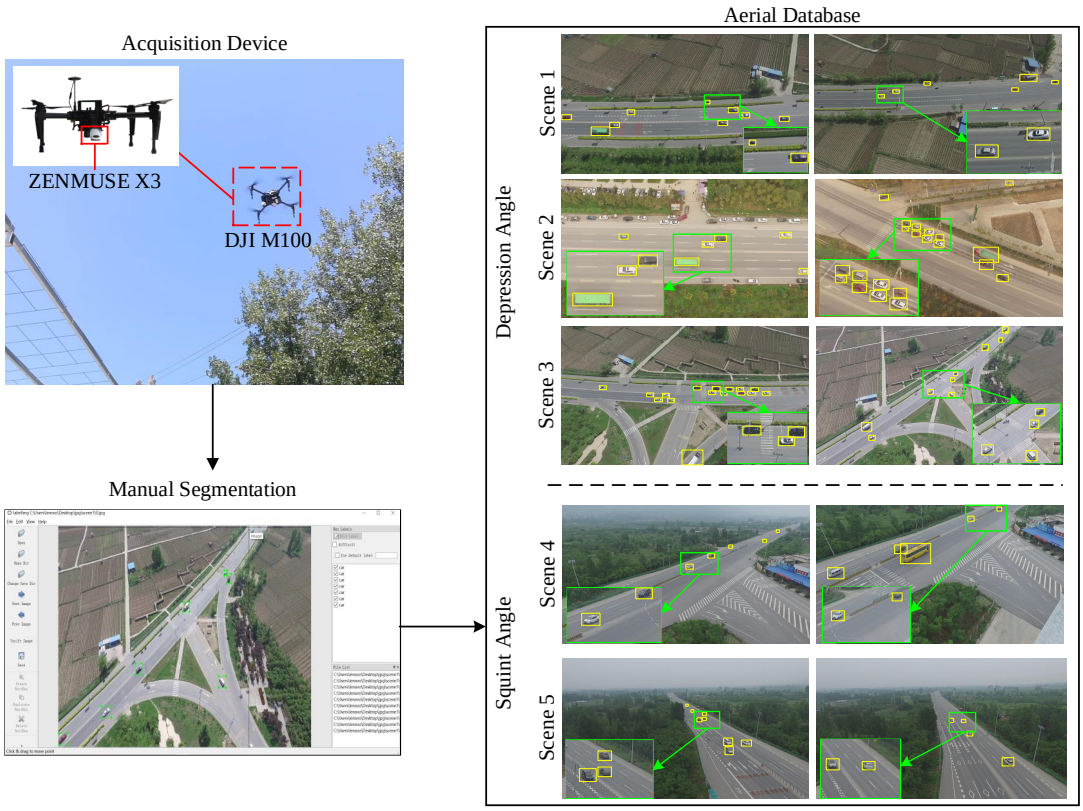

Figure 7. Experimental installation of aerial database. All test data were collected by DJI M100 with a ZENMUSE X3 camera (Shenzhen, China). We manually segmented its ground truth with the LabelImg tool. The right part of the figure shows some typical frames of the aerial video database. The yellow box is the ground truth of each frame.

3.2. Foreground Segmentation Result and Analysis

Multi-cue foreground segmentation aims to extract the potential moving target that needs visual detail augmentation. We employed the combination of optical flow and the background modeling method as our foreground segmentation algorithm. The choice of these two methods aimed to overcome the drawbacks of single methodology and to obtain a more complete foreground segmentation effect from input aerial images. In this section, we describe the analysis of the advantages of the combination of the two methods from two aspects: algorithm principles and experiment results.

In terms of the algorithm principles, optical flow and background modeling extract the potential foreground moving region on the basis of different perspectives. Optical flow describes object motion information by using the temporal variation of pixels in image sequences and the correlations between adjacent frames. Background modeling detects the foreground target area by modeling the image background and comparing the differences between the current image and the background model. Background modeling detects the foreground target area by modeling the image background and comparing the differences between the current image and the background model. This method is sensitive to the image gray level change information, and it can capture small and weak changes in the image which meet the demands of small aerial target detection. However, holes exist on the object when the gray scale of the target is similar to the background. Compared with the background modeling algorithm, optical flow can maintain good performance in that situation. The reason for this situation is that optical flow pays attention to the motion information of input aerial images, and similar backgrounds also have a certain degree of motion. However, the optical flow method is insensitive to weak motion, which is advantageous for the background modeling method. Thus,

these two methods have a complementary effect. In addition, if this problem is investigated from another perspective, the purpose of foreground segmentation is to enhance the target saliency relative to the background. Background modeling and optical flow start from two different views to describe object characteristics. The combination of the two methods actually improves object saliency and that helps to segment potential object regions from the complex background. Therefore, background modeling combined with optical flow can provide complementary advantages and obtain more a better segmentation result for small target detection.

As for the foreground segmentation result, we compared the proposed method with the background modeling method Fast-MCD and the optical flow method Farneback in our test database. This was done to further prove that the combination of the two algorithms improves the effectiveness of the segmentation result. The experimental results are shown in the Figure 8. It can be seen from the figure that the two methods have their own advantages and disadvantages for small target detection. The background modeling method Fast-MCD was generally able to segment the moving targets in the scene by detecting the grayscale changes of the images, even though the target is very small, such as in Frame 2317. However, the foreground segmentation results obtained by this method were incomplete with many holes, and some targets were split into many blocks, such as in Frame 0009. On the contrary, the motion segmentation results obtained by the Farneback optical flow method were generally relatively complete, such as in Frame 2317. However, this method had difficulty detecting the distant small moving targets in the scene, such as in Frame 0009. Instead of using these two methods alone, we combined them to improve the effect of foreground segmentation and were able to extract more complete moving targets, even distant small targets.

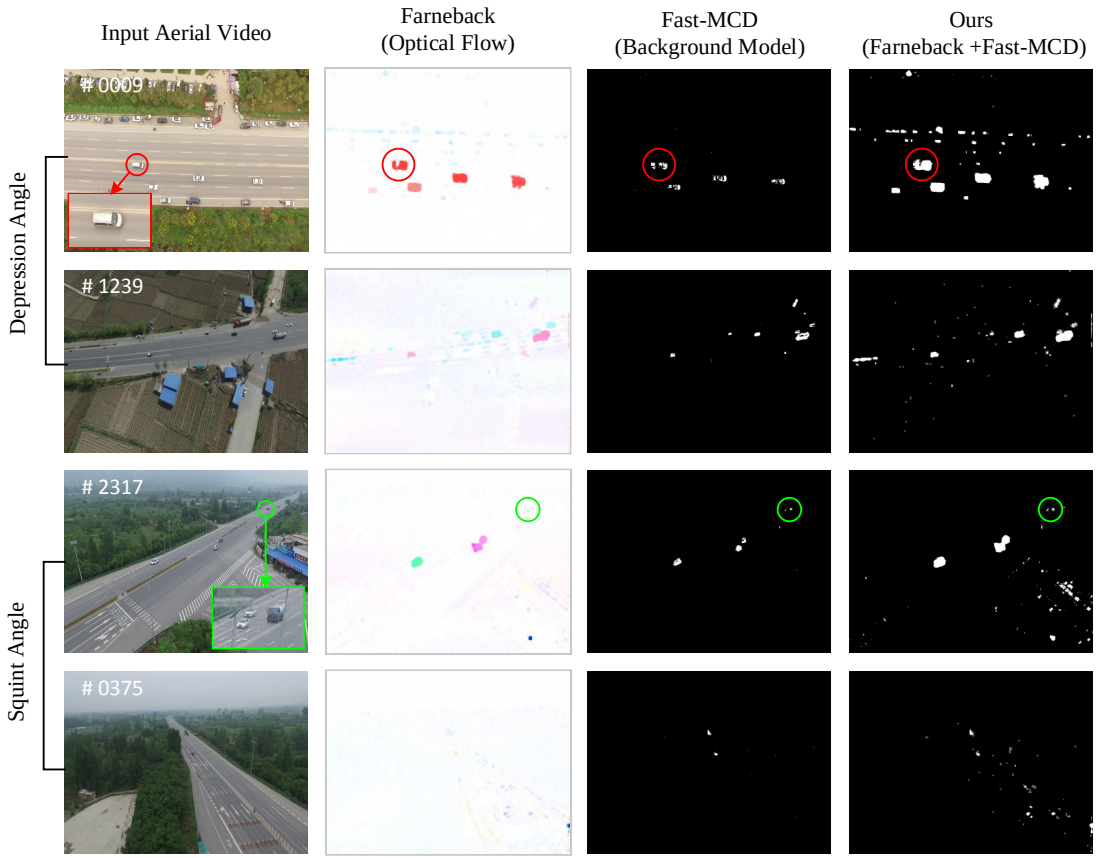

Figure 8. Contrast foreground segmentation result of the background modeling method Fast-MCD (the second column), the optical flow method Farneback (the third column), and the proposed multi-cue foreground segmentation method (the fourth column).

3.3. System Performance Evaluation

In this work, the performance of our system was evaluated on different aerial traffic scenes, including intersection and straight road, high loaded street and light loaded street, etc. Based on the shooting angle, we divided the detection results into two parts: the depression angle and the squint angle. Figure 9 displays some detection results, and Table 1 summarizes the detection performance of the quantitative analysis.

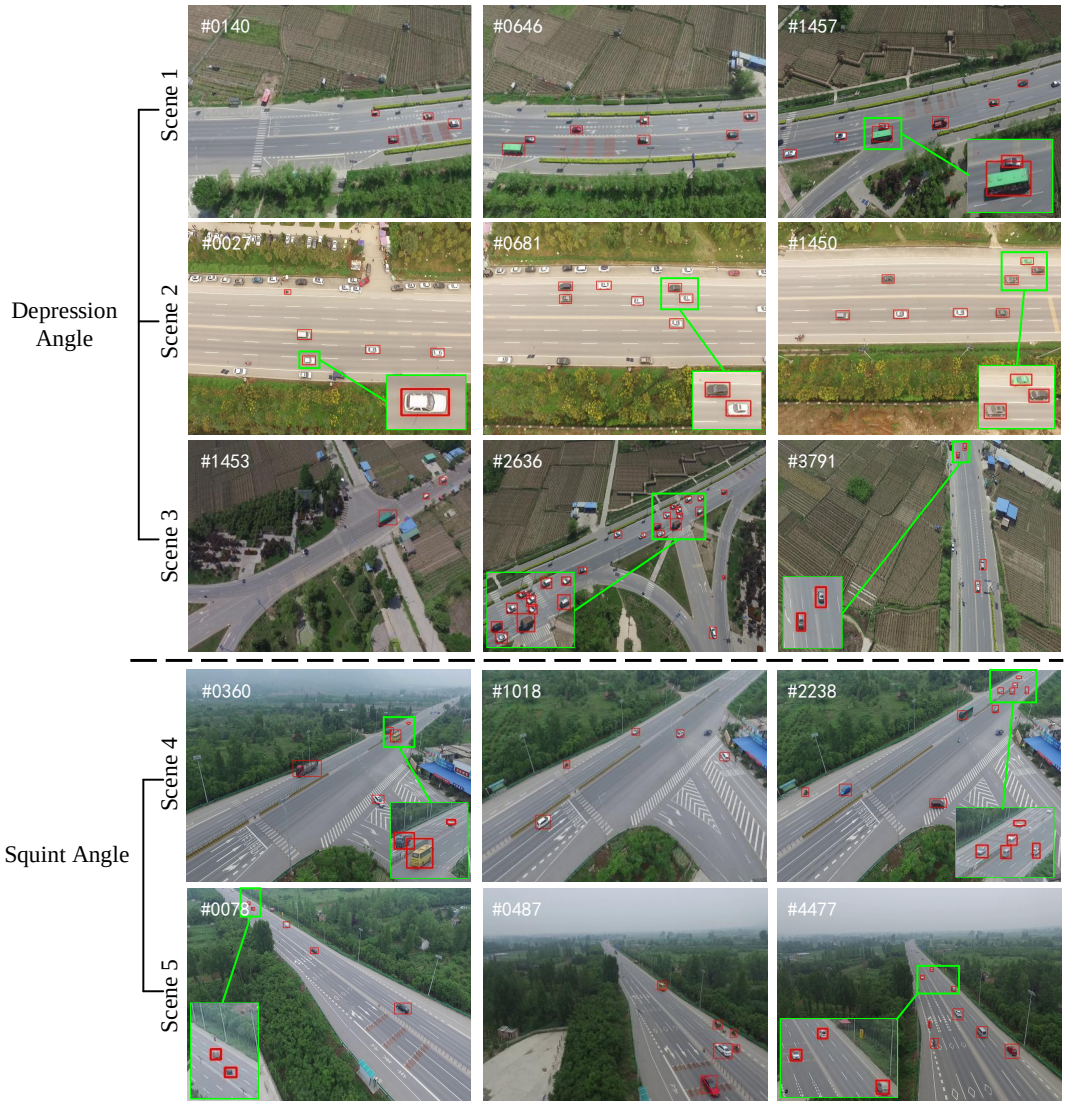

Figure 9. Some examples of small object detection results. The proposed method was evaluated in terms of the depression angle and splint angle. The number in the upper left corner is the frame number of each scenario.

In Figure 9, the red rectangles are the moving objects detected by the proposed method, and our system effectively detected moving targets from complex traffic scenarios. Next, we analyzed the detection results of the two shooting angles in detail. For the depression angle, there was little difference between the aerial object size under the same flying height. The average occupation ratio of each object in Scene 1 was lower than 1.0%, but the proposed method was able to accurately frame the interested small targets. Remarkably, when there was adhesion between the green bus and little car in Frame 1457, our system boxed them off well. Scene 2 was a high loaded street near a parking lot, and almost all moving cars were completely detected. From the enlarged details on the lower right, we can

see that the bounding boxes of each car were especially appropriate. Similarly, our system worked well in Scene 3, which was a busy traffic crossing. Frame 2636 contained a lot of targets with small sizes and intervals. For this case, we still obtained an encouraging detection result. As for the squint angle, the distance to the video camera caused the diversity of the target size. In the test database, the minimum object size was smaller than 10×10 and the maximum size was larger than 100×100. However, the proposed method showed great detection performance for both the intersection (as Scene 4 shown) and the straight road (as shown in Scene 5). In Frame 2238 in Scene 4, there were several tiny cars that are difficult to distinguish by human beings, and our method was able to detect them precisely. Our approach not only deals well with small, far-away targets but can also cover different object scales. The target size of Frame 4477 in Scene 5 had great variation and its detection result indicates that the detection system performed robustly in this situation. All of the above results demonstrate the great detection performance of our method.

In order to further evaluate the proposed method, we then carried out a quantitative analysis as shown in Table 1. We calculated the performance indexes of each scene separately, including the Precision(↑), Recall (↑), F1-score (↑), and Intersection-over-Union (IOU) (↑). The F1-score is a weighted harmonic mean of precision and recall and is a comprehensive evaluation parameter. IOU is a metric to measure the location accuracy by calculating the coincidence degree between the detected bounding box and the ground truth. If the IOU is greater than 0.5, this paper considered this object to be detected. The recall of each scene was more than 70% and the precision was more than 90%. Thus, as a comprehensive measure of detection, the F1-Score of our system was satisfied. As for the IOU, we determined that the position of the detection bounding box in each scene was very close to the ground truth, and the IOU value was 0.88 on average.

Table 1. Detection performance of the proposed method.

Angle	Scene	Precision	Recall	F1-Score	IOU
Depression Angle	**Scene1**	87.57%	82.26%	0.8483	0.8598
	Scene2	96.91%	85.22%	0.9069	0.8426
	Scene3	96.54%	92.24%	0.9434	0.9531
Squint Angle	**Scene4**	94.91%	76.48%	0.8470	0.8593
	Scene5	93.08%	81.43%	0.8687	0.9128

That means the position error was very small. In addition, we found that the recall of aerial videos in terms of the squint angle was a little lower than that of the depression angle. This is because the aerial video in the squint angle contains more different object sizes and that brings more difficulties for detection.

To conclude, extensive experiments indicate the effectiveness of the proposed method in both recall and precision of detection. On the one hand, almost all small moving targets can be accurately detected from the input aerial video. On the other hand, for different target sizes, the proposed approach can realize good detection performance at the same time. The great performance can satisfy users' demands well on small moving aerial object detection.

4. Discussion

To further validate our approach, we comprehensively compare it with other target detection works in this section. First of all, the comparison between YOLO v2 and the proposed method is presented to confirm the algorithm's effectiveness. In addition, we also provide contrast experiments with the other three methods for sufficiency and objectivity.

4.1. Comparison with the YOLO v2 Deep Detection Network

Since our system was proposed on the basis of the YOLO v2 deep detection network, we first compare our performance with it. In this section, considering that the detection result is greatly

influenced by the detection model, we employed the same model which was pre-trained by a set of normal size objects to make it fair. Figure 10 and Table 2 provide the comparison of our approach with YOLO v2 in terms of the average recall and accuracy on small vehicle detection.

Figure 10. Some detection results of the contrast methods in three scenarios, including (**a**) a traffic crossing, (**b**) T-junction and (**c**) busy trafficway. The rectangles in the lower left corner are the enlarged details. As we can see, the detection results of the proposed method were better than those of the comparison methods in terms of completeness and accuracy.

Figure 10 displays three typical examples of the two methods' results. As the figure shows, the first row is the ground truth of each scene. Scene (a) is a traffic crossing with a complex background and a large number of targets. In this case, the proposed method was able to detect almost all moving objects well, while the YOLO v2 network contained false alarms and some objects close to each other were mistakenly detected together. The problems with YOLO v2 not only existed in the crowded scene but also in the non-congested road. False alarms also appeared in Scene (b) with smooth traffic. What is worse, the YOLO v2 network missed almost half of the small targets placed far from the camera. In terms of the accuracy of the bounding box, Scene (c) and the boxes of our system were obviously more accurate than those detected by the YOLO v2 network. Therefore, the proposed method obtained a better qualitative performance. Next, we conducted a qualitative comparison between the two methods, as shown in Table 2. Both the precision rate and the recall rate of the proposed method were higher than those of YOLO v2, and the F1-Score increased by approximately 0.1. IOU was used to quantitatively indicate the detection position error, and the proposed method achieved a higher IOU value.

Table 2. Comparison detection performance with YOLO v2.

Angle	Method	Precision	Recall	F1-Score	IOU
Depression	**YOLO v2 [24]**	70.98%	77.87%	0.8141	0.7441
	YOLO v2 + Ours	**93.67%**	**86.57%**	**0.8998**	**0.8852**
Squint Angle	**YOLO v2 [24]**	56.18%	64.81%	0.6019	0.7205
	YOLO v2 + Ours	**94.29%**	**83.57%**	**0.8861**	**0.8861**

In addition, we also studied the performances of the two methods with objects of different sizes. Specifically, we divided the objects into four ranges: [100 × 100, 75 × 75), [75 × 75, 50 × 50), [50 × 50, 25 × 25), and smaller than 25 × 25. Then, we calculated the target detection precision in each range, as shown in Figure 11. When the object size was large, the two methods performed similarly. However, with size reduction, the proposed method was able to maintain greater precision on objects smaller than 25 × 25. The smaller the object size was, the larger detection precision gap between the two methods was. This further confirms the effectiveness of our system for small target detection. The proposed method improves the detection performance while also expanding the detectable object size range.

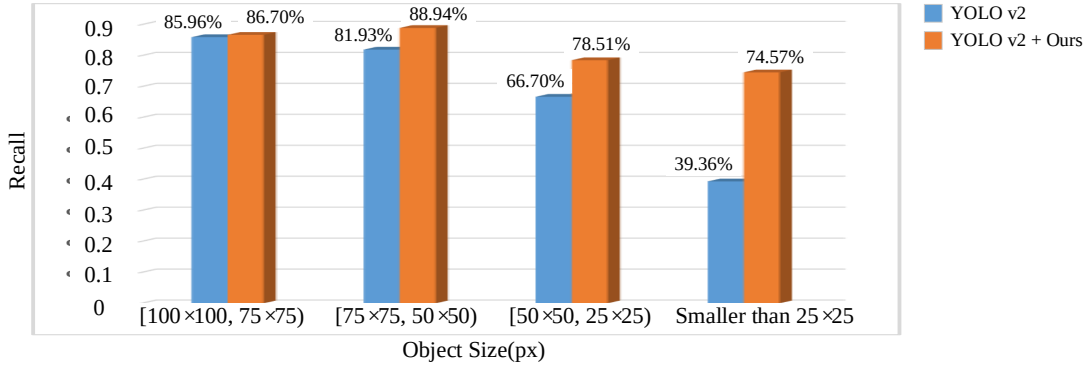

Figure 11. Detection performance with YOLO v2 on different object sizes. With a reduction in target size, the proposed method showed a huge advantage in terms of small target detection as it maintained great precision on objects smaller than 25 × 25.

The results above prove the strength of the visual detail augmented mapping approach for small aerial target detection. At the same time, its efficiency was also shown to be satisfactory. According to experimental testing, the average computing time was 77.63 ms which is sufficient for use in real-time, and the YOLO v2 network worked faster, around 31.25 ms. With this small increase in the amount of

computation, the proposed algorithm greatly improved the detection performance compared with direct deep network detection.

4.2. Comparison with State-of-the-Art Methods

In order to further explore the detection performance of the proposed method, we chose three classic and state-of-the-art methods for contrast experiments.

(1) Fast-MCD: Fast-MCD, which was proposed by Yi et al. in 2013, has been one of the most widely-used background subtraction algorithms in the field of target detection. The background is modeled through the dual-mode Single Gauss Model (SGM) with age, and camera motion information is obtained by the mixture of neighboring models. Fast-MCD is sensitive to subtle changes, and it helps to detect small aerial objects more accurately.

(2) RSS [27]: By considering both the stability and the saliency of small targets, this algorithm presents a novel model, called "RSS". RSS combines the regional stability and saliency to help with figure-ground segregation. It integrates the stability and saliency maps in a pixel-wise multiplication manner to remove false alarms. Experimental results have shown that this model adapts to target size variations and performs favorably in terms of both precision and recall.

(3) YOLO v3 [28]: This algorithm contains some updates to YOLO and includes a bunch of little design changes to achieve better detection performance. Compared with YOLO v2, YOLO v3 employs logistic regression to predict the objectness score for each bounding box, extracts features from three different scales using a similar concept to feature pyramid networks, and increases the number of anchor and convolutional layers. It has the advantages of a low false background detection rate and great performance on small target detection by changing the network structure.

In the specific implementation of these experiments, the Fast-MCD, RSS, and YOLO v3 codes were obtained from open source code. The parameters used in these methods were the original settings. For the evaluation standard, we used Precision (↑), Recall (↑), F1-score (↑) and IOU (↑) to judge performance in small moving object detection.

Figure 10 shows a comparison of the performance of the proposed method and the other approaches. As the figure shows, Fast-MCD had a high recall rate in our experiment, but its high false alarm rate had negative influences on detection performance. That means that the Fast-MCD, which is effective for obtaining motion information, cannot detect small moving objects from an aerial image well. RSS also wrongly framed some background regions as small objects, and there were still many missed detection cases when this method was used. Therefore, this method also cannot solve the problem of small aerial object detection well. YOLO v3 showed a low background false detection rate, but it missed some real targets. However, with the proposed method, almost all objects with different sizes were accurately detected from the input aerial video. Moreover, in the third column in Figure 10, the bounding box of our method is shown to be the most suitable among the comparison methods. The proposed method based on visual detail augmented mapping showed obvious advantages, not only quantitatively but also qualitatively. Table 3 provides the quantitative comparison results, and our system outperformed the other methods in terms of precision, recall, F1-Score, and IOU. The precision and recall rate of our method were more than 80% for every shooting angle. Through calculations, its F1-Score was greater than 0.85 which was the best performance among all contrastive experiments. Meanwhile, the IOU of the proposed method was more than 0.88, which suggests that accurate object positioning occurred. Remarkably, the precision of YOLO v3 in terms of the depression angle and the recall rate of Fast-MCD in terms of the squint angle were higher than the recall and recision of the proposed method, respectively. However, the other measures of the two methods were 77.12% and 65.80%, respectively, which means that there was a high missed detection rate and false alarm rate. Moreover, we found that these algorithms had lower performance with the squint angle than with the depression angle. This is because the aerial video in the squint angle contains different moving object sizes ($150 \times 150 - 10 \times 10$), and the target detection is more difficult. We can see that F1-Score of our algorithm reduced from 0.8998 to 0.8861, and the F1-Score

of YOLO v3 reduced from 0.8587 to 0.6699, which shows that our algorithm is more effective and robust for small moving target detection. Though our method had a slightly lesser performance on one measure, its comprehensive performance in terms of the F1-Score was the highest. This indicates that the proposed method is sufficient to meet the demands of small moving object detection from aerial data. In terms of computational efficiency, YOLO v3 was shown to be the fastest with a computing time of about 46.51 ms. With visual detail augmented mapping, the average detection time of our system was 77.63 ms. Though the proposed method is slower than other methods, it also can satisfy the real-time needs of detection system. By synthesizing the two aspects of efficiency and precision, our approach performs better in small target detection than the other methods.

Table 3. Comparison of the detection performance of various methods in terms of Precision, Recall, F1-Score and IOU.

Angle	Method	Precision	Recall	F1-Score	IOU
Depression Angle	**Fast-MCD [23]**	62.63%	82.93%	0.7136	0.7669
	RSS [27]	31.68%	25.44%	0.2822	0.6520
	YOLO v2 [24]	70.98%	77.87%	0.8141	0.7441
	YOLO v3 [28]	**96.85%**	77.12%	0.8587	0.7649
	YOLO v2 + Ours	93.67%	**86.57%**	**0.8998**	**0.8852**
Squint Angle	**Fast-MCD [23]**	65.80%	**94.75%**	0.7767	0.6925
	RSS [27]	29.95%	20.16%	0.2410	0.5947
	YOLO v2 [24]	56.18%	64.81%	0.6019	0.7205
	YOLO v3 [28]	94.13%	52.00%	0.6699	0.7682
	YOLO v2 + Ours	**94.29%**	83.57%	**0.8861**	**0.8861**

The reasons for the high performance of our system are as follows. First, Fast-MCD only considers information about the motion of the moving object. This information is susceptible to noise interference and unrelated things. Thus, this causes the high false alarm rate of Fast-MCD. Second, RSS is based on the stability and saliency of small objects, and this approach is suitable for single small target detection under less complex backgrounds. However, the views of aerial visual data are usually large and their backgrounds are usually complicated with a large amount of interference. This leads to difficulty for RSS in detecting moving objects from aerial video. Third, YOLO v3, which achieves a comparable performance to our method, is based on the use of a number of convolution layers to extract deep features for better object representation. Compared with YOLO v3, which improves the detection rate of small targets by improving the YOLO v2 network structure, our algorithm greatly improves the detection rate of small moving targets by using a visual detail augmented mapping method combined with the YOLO v2 network, which can achieve better detection results than YOLO v3.

5. Conclusions

In this paper, we proposed a novel visual detail augmented mapping approach for small aerial moving target detection. To address this, we first presented a multi-cue foreground segmentation method which combines the optical flow algorithm Farneback and the background modeling method Fast-MCD to obtain motion and grayscale change information of input aerial video. Through comprehensive analysis of the two clues, the potential moving target regions are extracted out. Then, a visual detail augmented mapping approach was proposed which maps the initial aerial image to a new foreground space. This mapping consists of two modules: one is multi-resolution mapping, which provides more detailed target information for the subsequent detection network; the other is foreground augmented mapping, in which the original potential moving target regions are mapped to a more valuable foreground augmented map. The new foreground augmented map is a visual augmentation of the original aerial image. Finally, driven by the YOLO v2 deep detection network and the coordinate inverse calculation, a small moving target detection system is implemented.

In order to evaluate the performance of the system, we carried out a lot of experiments. Through the analysis of foreground segmentation experiment results, it was proved that the algorithm can greatly improve the integrity of foreground segmentation results by combining the background modeling algorithm and the optical flow method. On this basis, extensive experimental results were used to demonstrate that the proposed method is efficient and robust for small moving target detection without changing and retraining the deep detection network. In addition, the comparative experiments with the recently-published state-of-the-art methods show that our system performs the best in terms of both detection speed and accuracy. Our future work will focus on developing object tracking and re-identification based on the proposed method.

Author Contributions: J.L., Y.D., C.L. and T.Y. designed the overall system and wrote the source code. J.L., Y.D., C.L., J.S. and D.L. participated in the research data collection, analysis and interpretation. T.Y. guided the experiments and the statistical analysis, and Z.L. supplied help with the experiments and paper revision. Additionally, they jointly designed and performed the experiments.

Funding: This research was funded by the National Natural Science Foundation of China [No. 61502364, No. 61672429, No. 61272288].

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

References

1. Chen, T.; Pennisi, A.; Li, Z.; Zhang, Y.; Sahli, H. A Hierarchical Association Framework for Multi-Object Tracking in Airborne Videos. *Remote Sens.* **2018**, *10*, 1347. [CrossRef]
2. Kondo, M.; Shoji, R.; Miyake, K.; Furuya, T.; Ohshima, K.; Shimizu, E.; Inaishi, M.; Nakagawa, M. Monitor System for Remotely Small Vessel Navigating. In Proceedings of the International Conference on Human Interface and the Management of Information, Las Vegas, NV, USA, 15–20 July 2018; pp. 419–428.
3. Liu, X.; Yang, T.; Li, J. Real-Time Ground Vehicle Detection in Aerial Infrared Imagery Based on Convolutional Neural Network. *Electronics* **2018**, *7*, 78. [CrossRef]
4. Yang, X.; Sun, H.; Fu, K.; Yang, J.; Sun, X.; Yan, M.; Guo, Z. Automatic Ship Detection in Remote Sensing Images from Google Earth of Complex Scenes Based on Multiscale Rotation Dense Feature Pyramid Networks. *Remote Sens.* **2018**, *10*, 132. [CrossRef]
5. Risse, B.; Mangan, M.; Webb, B.; Pero, L.D. Visual Tracking of Small Animals in Cluttered Natural Environments Using a Freely Moving Camera. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 2840–2849.
6. Li, J.; Li, C.; Yang, T.; Lu, Z. A Novel Visual Vocabulary Translator Based Cross-Domain Image Matching. *IEEE Access* **2017**, *5*, 23190–23203. [CrossRef]
7. Wang, G.; Wang, X.; Fan, B.; Pan, C. Feature Extraction by Rotation-Invariant Matrix Representation for Object Detection in Aerial Image. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 851–855. [CrossRef]
8. Yang, T.; Li, J.; Yu, J.; Wang, S.; Zhang, Y. Diverse Scene Stitching from a Large-Scale Aerial Video Dataset. *Remote Sens.* **2015**, *7*, 6932–6949. [CrossRef]
9. Li, J.; Liang, X.; Wei, Y.; Xu, T.; Feng, J.; Yan, S. Perceptual Generative Adversarial Networks for Small Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1951–1959.
10. Yang, T.; Wang, X.; Yao, B.; Li, J.; Zhang, Y.; He, Z.; Duan, W. Small Moving Vehicle Detection in a Satellite Video of an Urban Area. *Sensors* **2016**, *16*, 1528. [CrossRef]
11. Chen, X.; Kundu, K.; Zhu, Y.; Ma, H.; Fidler, S.; Urtasun, R. 3D Object Proposals Using Stereo Imagery for Accurate Object Class Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1259–1272. [CrossRef]
12. Cao, G.; Xie, X.; Yang, W.; Liao, Q.; Shi, G.; Wu, J. Feature-Fused SSD: Fast Detection for Small Objects. *arXiv* **2017**. arXiv:1709.05054v2.
13. Li, H.; Lin, Z.; Shen, X.; Brandt, J.; Hua, G. A Convolutional Neural Network Cascade for Face Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5325–5334.

14. Cai, Z.; Fan, Q.; Feris, R.S.; Vasconcelos, N. A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 354–370.
15. Viola, P.A.; Jones, M.J. Rapid Object Detection using a Boosted Cascade of Simple Features. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, HI, USA, 8–14 December 2001; pp. 511–518.
16. Girshick, R.B.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
17. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1440–1448.
18. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef]
19. Chen, C.; Liu, M.; Tuzel, O.; Xiao, J. R-CNN for Small Object Detection. In Proceedings of the Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; pp. 214–230.
20. Yang, F.; Choi, W.; Lin, Y. Exploit All the Layers: Fast and Accurate CNN Object Detector with Scale Dependent Pooling and Cascaded Rejection Classifiers. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1July 2016; pp. 2129–2137.
21. Ren, Y.; Zhu, C.; Xiao, S. Small Object Detection in Optical Remote Sensing Images via Modified Faster R-CNN. *Appl. Sci.* **2018**, *8*, 813–824. [CrossRef]
22. Farneback, G. Two-Frame Motion Estimation Based on Polynomial Expansion. In Proceedings of the Scandinavian Conference on Image Analysis, Halmstad, Sweden, 29 June–2 July 2003; pp. 363–370.
23. Yi, K.M.; Yun, K.; Kim, S.W.; Chang, H.J.; Choi, J.Y. Detection of Moving Objects with Non-stationary Cameras in 5.8ms: Bringing Motion Detection to Your Mobile Device. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, USA, 23–28 June 2013; pp. 27–34.
24. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
25. UA-DETRAC. Available online: http://detrac-db.rit.albany.edu/ (accessed on 1 December 2017).
26. Tzutalin. LabelImg. Available online: https://github.com/tzutalin/labelImg (accessed on 1 September 2017).
27. Lou, J.; Zhu, W.; Wang, H.; Ren, M. Small Target Detection Combining Regional Stability and Saliency in a Color Image. *Multimed. Tools Appl.* **2017**, *76*, 14781–14798. [CrossRef]
28. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**. arXiv:1804.02767v1