

Air-to-ground multimodal object detection algorithm based on feature association learning

Dongfang Yang¹, Xing Liu² , Hao He² and Yongfei Li²

Abstract

Detecting objects on unmanned aerial vehicles is a hard task, due to the long visual distance and the subsequent small size and lack of view. Besides, the traditional ground observation manners based on visible light camera are sensitive to brightness. This article aims to improve the target detection accuracy in various weather conditions, by using both visible light camera and infrared camera simultaneously. In this article, an association network of multimodal feature maps on the same scene is used to design an object detection algorithm, which is the so-called feature association learning method. In addition, this article collects a new cross-modal detection data set and proposes a cross-modal object detection algorithm based on visible light and infrared observations. The experimental results show that the algorithm improves the detection accuracy of small objects in the air-to-ground view. The multimodal joint detection network can overcome the influence of illumination in different weather conditions, which provides a new detection means and ideas for the space-based unmanned platform to the small object detection task.

Keywords

Feature association, multimodal learning, air-to-ground detection, deep learning

Date received: 1 September 2018; accepted: 2 March 2019

Topic: Vision Systems

Topic Editor: Henry Leung

Associate Editor: Huaping Liu

Introduction

The object detection technology under the air-to-ground field of view is a special but widely used application of multi-object detection technology. In air-to-ground applications, the scale of view is large, and therefore the target of interest tends to be rather small, which will bring new difficulties in object detection.^{1–4} In the military field, object detection can be directly applied to tasks such as battlefield investigation, situation analysis, air-to-ground target strike, and object tracking. In other civilian fields, air-to-ground object detection can be directly applied to various tasks, such as traffic monitoring, natural disaster analysis, and agricultural ecological management. The object detection in the air-to-ground scenario has the following characteristics.

In air-to-ground applications, the visual distance is always long, in which the target seems to be rather small, and less feature information can be involved. In air-to-ground views, targets tend to have only a few tens of pixels of information, and convolutional neural networks (CNN) have much less information available in feature extraction than in conventional life scenarios. At the same time, the air-to-ground target scene has a large field of view, and the

¹ Xi'an University of Posts and Telecommunications, Xi'an, Shaanxi, China

² Xi'an Research Institute of High Technology, Xi'an, Shaanxi, China

Corresponding author:

Dongfang Yang, Xi'an University of Posts and Telecommunications, Xi'an, Shaanxi 710121, China.

Email: yangdf301@163.com



Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License

(<http://www.creativecommons.org/licenses/by/4.0/>) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

detection target has many environmental changes (occlusion, background interference), so it is hard to acquire satisfied detection results.

In addition, in the dark environment, at night for example, the object contour is not clear, and the feature information is lost. The traditional single-mode (visible light) detection poses a severe challenge, which makes the detection task unable to adapt to different lighting conditions. With the increasing research on air-to-ground object detection technology based on deep learning, the target detection method seems to be maturing. However, as the scene and application environment become more complex, the traditional single-mode visible light detection problems are gradually becoming prominent.

Most of the early object detection is concentrated in the framework of traditional computer vision and image processing. In order to achieve the feature description of the object, scale invariant feature transform (SIFT),¹ histogram of oriented gradient (HOG),² speeded up robust features (SURF),³ and other artificial design features play an important role. However, all the artificially designed features are difficult to obtain satisfactory results in complex background or subtle object detection and recognition applications due to their low dimension and insufficient target description. In recent years, the development of deep learning has provided us with an effective way of image description. With the advent of big data and the continuous development of high performance computing hardware, data-driven deep CNN have made great progress in feature extraction performance. In just a few years of development, feature extraction networks such as AlexNet,⁴ VGGNet,⁵ GoogLeNet,⁶ ResNet,⁷ and MobileNets⁸ have been proposed, which laid the foundation for the goal detection task using deep learning. Based on the feature extraction network, the development of object detection technology based on deep learning has gone through two stages: the two-stage detection algorithm based on region proposal and the one-stage detection algorithm based on regression. The previous stage is represented by RCNN.^{9–11} Among them, Faster RCNN¹¹ improves the feature extraction mechanism of Fast RCNN based on RCNN and combines multitask loss optimization with regional proposal network which achieves an end-to-end detection network. It can achieve real-time performance of object detection better than other RCNNs. In order to further improve the real-time performance of object detection, the You Only Look Once (YOLO)¹² algorithm is proposed outside the RCNN framework. This method abandons the process of Faster RCNN regional proposal and performs regression calculation on the randomly generated bounding box. This end-to-end detection further improves the real-time performance of the object detection algorithm. However, due to the constraints of its network structure, the typical YOLO algorithm encounters difficulties in performing small object problems. The emergence of Single Shot Multi-Box Detector (SSD)¹³ algorithm overcomes the inherent defects of

YOLO algorithm and effectively improves the performance of small object detection. A typical SSD detection network can be divided into two parts: a feature extraction subnetwork in the front end and a detection subnetwork in the back end. On the detection subnetwork side, the SSD combines the anchor idea of the Faster RCNN with the regression idea of YOLO, which generates a priori frames on six different scale feature maps for prediction, thus enriching the feature scale of the detection. Different feature maps have different receptive fields after convolution operations. Feature maps of different scales can predict boxes of more scales. However, due to lack of semantic information, shallow feature maps still have low detection performance, which limits its application.

In this article, a new network structure is designed based on the original SSD with the air-to-ground small object detection as the entry point. At the same time, based on the complementarity of infrared radiation (IR) and visible light images in practical applications, from the perspective of multimodality, the experimental study of multimodality detection under different illumination conditions is carried out. The main contributions of this article are:

1. In order to achieve the object detection task under different day and night conditions, this article integrates the long wave infrared and visible light data collected by the laboratory to produce a multimodal detection data set.
2. A new object detection model is proposed to correlate different receptive field feature map information to improve the accuracy.
3. Using the cross-modal conception, the IR image and the visible light image are jointly trained to realize the detection function under different illumination conditions. For example, under the condition of weak illumination at night, visible light can use less feature information, detection often fails, and temperature-sensing infrared image detection becomes more advantageous.

Related work

Using contextual relationships to assist object detection is a hot research direction of deep learning detection algorithms. For example, in the works of Chen et al.¹⁴ and Liu et al.,¹⁵ the ContextNet and ParseNet are proposed, respectively, which can improve the detection ability of small objects by merging the object feature map information and the lower level feature maps which have the object context information. Zhao et al.¹⁶ propose a PSPNet, which uses different pooling operations to generate feature maps of different receptive field sizes and then combines to increase the small object detection capability. In general, the

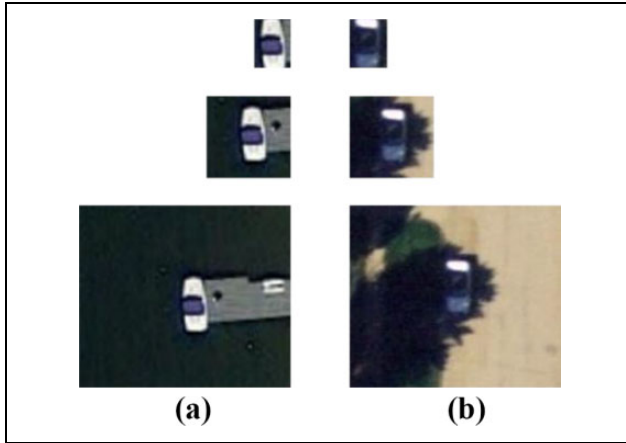


Figure 1. Different receptive field information in object detection.

mainstream method to improve the performance of small object detection is to generate feature maps of different size receptive fields and integrate context information to increase the accuracy of detection. For machine vision, the receptive field characterizes the range of perception of the image. As shown in Figure 1, when the target receptive field is small, it is affected by the imaging field of view, and less information is obtained by using the small receptive field, which is easy to cause misclassification (considered as a car). However, when expanding the receptive information and making full use of the target's more surrounding information and scene information, it is more effective to distinguish the real category of the object. In Figure 1(a), a boat parked by the lake and in Figure 1(b), a car on the road and it's shaded by trees.

The SSD model is a network that naturally uses different receptive field information formed by different scales to be detected. According to the limited calculation sources, we choose the faster SSD as the basic framework in our air-to-ground applications. The structure of SSD is shown in Figure 2.

The SSD network model consists of a feature extraction subnetwork and a detection subnetwork. The feature extraction subnetwork is usually a traditional convolutional network such as VGG16 and ResNet. A large number of improved models are proposed based on the unique structure of SSD network, such as FSSD¹⁷ which combines the feature extraction subnetwork low-level feature map and high-level feature map. And the fine-grained feature map is beneficial to the robustness of the detection algorithm. Benefiting from the unique structure of the SSD detection subnetwork, several feature fusion methods were proposed by using the context information. For example, RFBNet¹⁸ and RUN¹⁹ use the inception and ResNet network structures, respectively. They use the multibranch convolution to simulate the change of the receptive field and join the residual module to improve the ability of the feature extraction. Another version is the fusion of feature maps, such as

DSSD²⁰ and RON,²¹ which use the FPN²² framework to fuse different semantic information (as shown in Figure 3(a)). In addition, DSOD,²³ RRC,²⁴ and RSSD²⁵ fuse the deep feature map with the shallow feature map (as shown in Figure 3(b)) by using the deconvolution and pooling tricks. It enriches the fine-grained and topological information of different size feature maps.

In deep learning algorithm, it's highly depending on training data. The well-known Pascal VOC and Microsoft COCO can be used in object detection tasks, whereas in air-to-ground applications, data sets are affected by the brightness conditions, and the number of concerned objects is relatively few than that in VOC and COCO. Air-to-ground data sets are always based on visible light cameras. Since there are no official data sets in multiple modalities, it is impossible to realize multi-weather detection tasks in different environments, without specialized air-to-ground data sets. In artificial perception field, the multimodal fusion is considered as an effective way to enhance the information completeness of the environment.²⁶ The multimodal fusion attracted attentions in various applications, such as material retrieval,²⁷ image annotations,²⁸ and robotic perception.²⁹ Actually, the air-to-ground detection process is rather a similar process of robot perception, thus this article presents a novel object detection method by using multimodal observations.

Multimodal detection and feature association detection model

In poor brightness conditions, the use of temperature distribution to characterize the infrared image of an object has great advantages in target description. Therefore, it is urgent to establish a multimodal detection model compatible with visible and IR images. Whereas, infrared thermal image characterizes the temperature distribution of the scene and has no stereoscopic sense, so the resolution is low and the resolution potential is poor for human eyes. On the other hand, due to the thermal balance of the scene, long wavelength, long transmission distance, and atmospheric attenuation, the infrared image has strong spatial correlation, low contrast, and blurred visual effects. Therefore, using deep learning algorithm to describe IR image is a valuable work. The fusion of IR and visible light is visualized by using picture-in-picture method directly, and it can combine visible light and infrared information to enrich the purpose of the characterization. As shown in Figure 4, under daylight conditions, visible light data can clearly represent the contour and feature information of the target. Under this kind of circumstance, it is more suitable for visible light image. When the light is dark, visible and IR fused picture-in-picture data sets can compensate for the loss of characteristic information of visible light data set due to insufficient light. And when the nighttime illumination information is seriously insufficient, the visible light

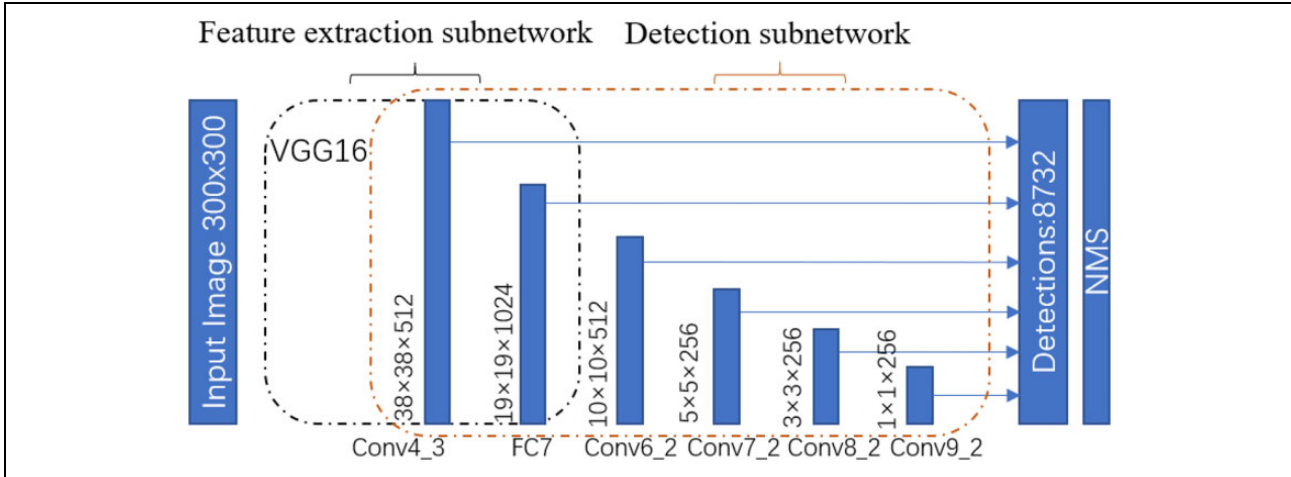


Figure 2. SSD structure. SSD: Single Shot Multi-Box Detector.

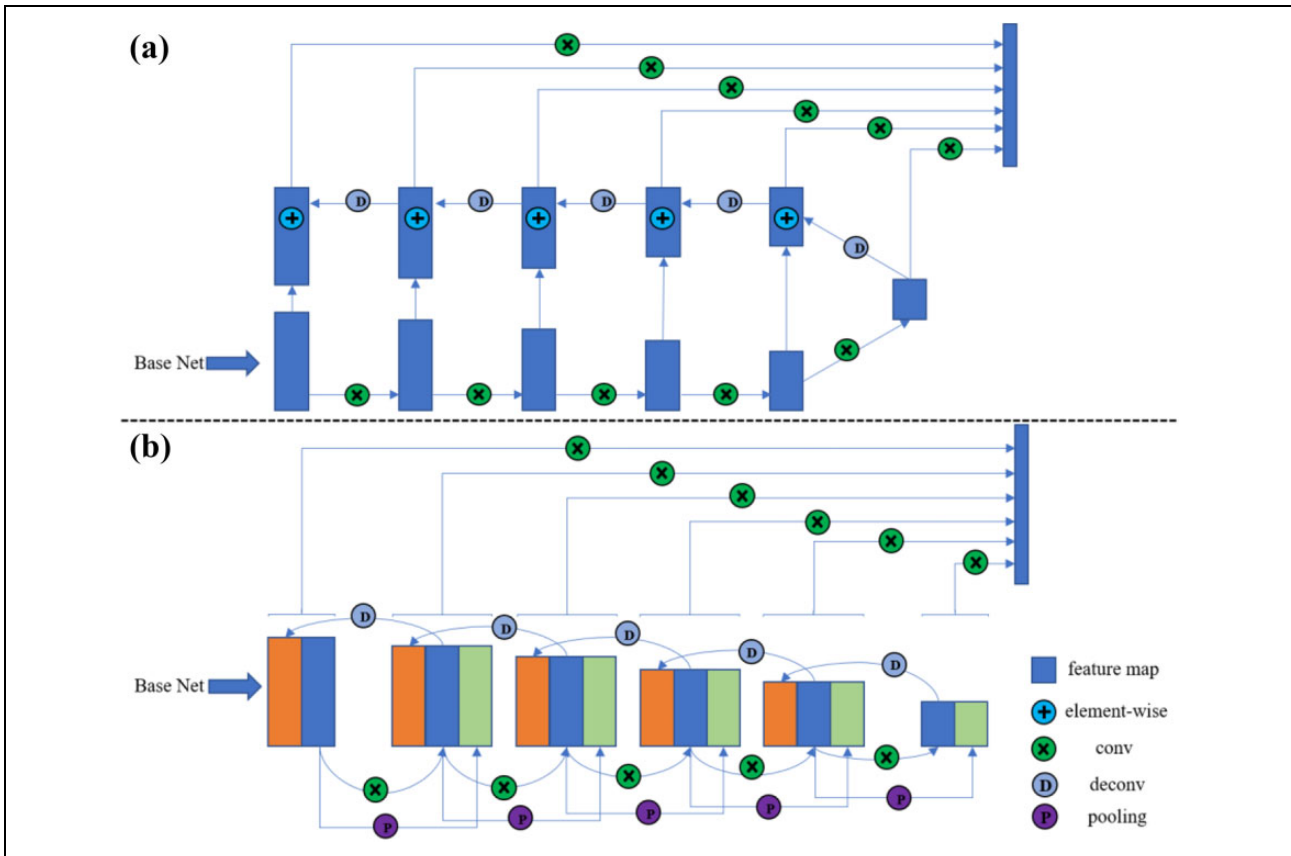


Figure 3. The mainstream way of SSD feature association. SSD: Single Shot Multi-Box Detector.

target almost disappears, and infrared imaging can play an important role.

Feature maps of different sizes represent different scale of receptive field, which can be utilized to build the contextual relationships. The general SSD object detection algorithm often learns the feature information from small receptive field to large receptive field by using multi-scale boxes in feature maps. But this class

of learning methods pays less attention to the context-related problem. Reasonable use of different sizes to feel the relationship between the fields is beneficial to the object detection process. In traditional methods, the simple fusion method makes the network itself lack the filtering control ability. In Faster RCNN, YOLO, and SSD, the posterior feature map learns the information from the front layers, which is derived from the

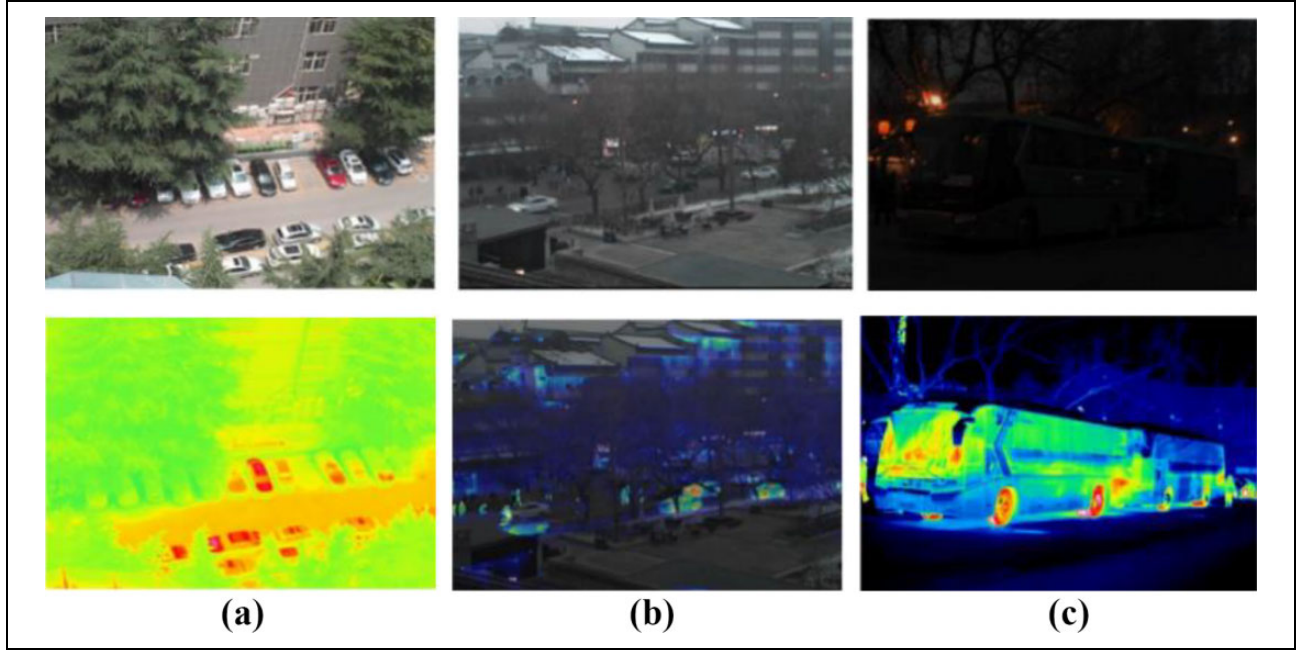


Figure 4. Comparison between white light images and IR images in different brightness environments. (a) Strong brightness conditions, (b) low brightness conditions, and (c) dark conditions.

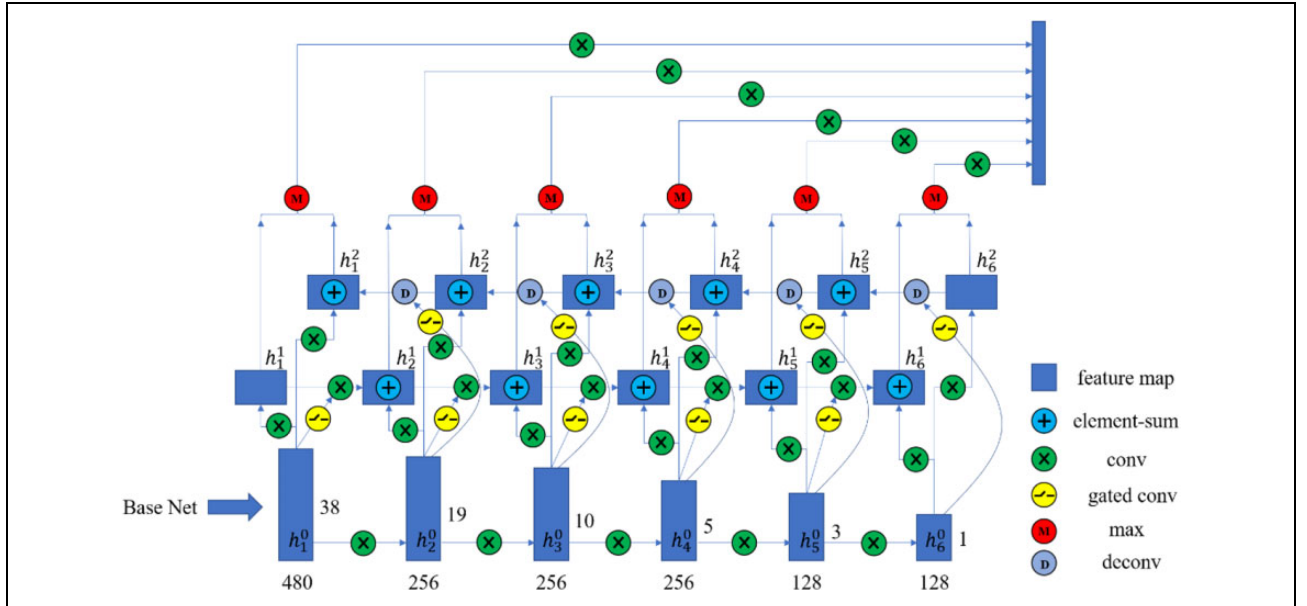


Figure 5. Feature association detection model.

feedforward nature mechanism of the neural network. However, it is obvious that the information of back layer feature map can also affect the front layer features. This bidirectional associative feature map can better learn the deep topological relationship between different feature maps. In this article, combined with the unique structural relationship of SSD series detection network, an object detection algorithm for feature association learning is proposed. The feature correlation detection network is shown in Figure 5.

The main structure uses a bidirectional feature information flow, and the core module is a gated convolution operation. The bidirectional feature information flow is divided into bottom-up and top-down parts. The bottom-up information flow (h_i^1) indicates the flow of information from the small to the large receptive field. In this process, the semantic information is gradually enhanced. In addition, the top-down information flow (h_i^2) represents the flow of information from the large to the small receptive field, and the semantic information is gradually weakened. This

bidirectional information flow is controlled by the gating function, and the deep topology association between different feature maps is adaptively learned through continuous training of the network. Among them, the way the information flow propagates from the bottom up is depicted in equation (1)

$$h_i^1 = \sigma(h_i^0 \otimes w_i^1 + b_i^{0,1}) + \sigma(h_{i-1}^0 \otimes w_{i-1,i}^1 + b_i^1) \quad (1)$$

The top-down propagation method of information flow is written in equation (2)

$$h_i^2 = \sigma(h_i^0 \otimes w_i^2 + b_i^{0,2}) + \sigma(h_{i+1}^2 \otimes w_{i,i+1}^2 + b_i^2) \quad (2)$$

The two types of information are integrated in equation (3)

$$h_i^3 = \sigma(\max(h_i^1, h_i^2) \otimes w_i^3 + b_i^3) \quad (3)$$

In the feature map h_i^k , there are multiple feature channels, so the feature map will inevitably have redundant information. Secondly, the transmission response of the information should be controlled by a filter that can correlate the information of different receptive fields. Therefore, the network implements the abovementioned operation by means of a gate function

$$h_i^1 = \sigma(h_i^0 \otimes w_i^1 + b_i^{0,1}) + G_i^1 \cdot \sigma(h_{i-1}^0 \otimes w_{i-1,i}^1 + b_i^1) \quad (4)$$

$$h_i^2 = \sigma(h_i^0 \otimes w_i^2 + b_i^{0,2}) + G_i^2 \cdot \sigma(h_{i+1}^2 \otimes w_{i,i+1}^2 + b_i^2) \quad (5)$$

$$G_i^1 = \text{sigm}(h_{i-1}^0 \otimes w_{i-1,i}^g + b_{i-1,i}^g) \quad (6)$$

$$G_i^2 = \text{sigm}(h_{i+1}^0 \otimes w_{i+1,i}^g + b_{i+1,i}^g) \quad (7)$$

In the above formulations, $\text{sigm}(x) = 1/[1 + \exp(-x)]$ is the sigmoid operation for each tensor value, \cdot is the product of each tensor value, \otimes is the convolution operation, and \otimes is the deconvolution operation.

The loss function of the model is defined as follows

$$L_{\text{dec}}(x, c, l, g) = \frac{1}{N} (L_{\text{conf}}(x, c) + \alpha L_{\text{loc}}(x, l, g)) \quad (8)$$

where N is the number of a priori boxes that match the default box. If $N = 0$, the loss is 0. c, g are the class label and box coordinate parameters of the ground truth. x, l are the predicted class labels and the box coordinate parameters. α is a trade-off parameter.

The optimization function of box regression loss is defined as follows

$$L_{\text{loc}}(x, l, g) = \sum_{i \in \text{Pos}} \sum_{m \in \{cx, cy, w, h\}} x_{ij}^k \text{smooth}_{L1}(\bar{l}_i^m - \bar{g}_j^m) \quad (9)$$

herein

$$\bar{g}_j^{cx} = \frac{g_j^{cx} - d_i^{cx}}{d_i^{cx}}, \bar{g}_j^{cy} = \frac{g_j^{cy} - d_i^{cy}}{d_i^{cy}}, \bar{g}_j^w = \log\left(\frac{g_j^w}{d_i^w}\right), \bar{g}_j^h = \log\left(\frac{g_j^h}{d_i^h}\right)$$

The optimization function of confidence loss uses the softmax and is defined as follows

$$L_{\text{conf}}(x, c) = - \sum_{i \in \text{Pos}} x_{ij}^p \log(\bar{c}_i^p) - \sum_{i \in \text{Neg}} \log(\bar{c}_i^0) \quad (10)$$

where $\bar{c}_i^p = \exp(c_i^p) / \sum_p \exp(c_i^p)$ is the category confidence, l is the prediction box parameter, g is the ground-truth box parameters. cx, cy, w, h is the coordinates and length and width of the prediction box, and x_{ij}^p is the indicator function (when the i th default box and the j th ground-truth box match with category c , the value is 1, otherwise 0).

Experiments

In this article, the multimodal data sets are collected from white light camera and long-wave IR camera simultaneously. The same scene was measured in three different styles, which include white light, IR, and white-IR fused mode, as shown in Figure 6. In order to validate the environmental adaptability of detection algorithm, different light conditions are involved in our data sets. In this experiment, we choose a car and a bus as the concerned targets. The total number of data sets contains 1500 images, including 500 visible light images, 500 infrared images, and 500 fused images.

After the multimode data sets were sampled, they are trained and tested under the framework of the proposed detection network. Ubuntu 16.04 (64-bit) and the PyTorch framework are involved, and the NVIDIA GTX 1080 Ti graphics card (11G) is utilized. In order to evaluate the performance of the proposed algorithm, the Pascal VOC2010 verification standard³⁰ is adopted in the experiments. Where the accuracy object type is verified by the average accuracy, it's defined as follows

$$f_{\text{AP}} = \int_0^1 P(R) dR \quad (11)$$

herein P is the accuracy and R is the recall rate. They are calculated as follows

$$P = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}} \quad (12)$$

$$R = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}} \quad (13)$$

where N_{TP} is the number of targets correctly identified, N_{FP} is the number of nontargets as targets, and N_{FN} represents the number of targets not identified. Therefore, $N_{\text{TP}} + N_{\text{FP}}$ represents the number of targets identified, and $N_{\text{TP}} + N_{\text{FN}}$ represents the total number of existed targets. On the basis of the aforementioned accuracy index, mean average precision (mAP) is used to measure the comprehensive performance of the detector for different categories, which is defined as

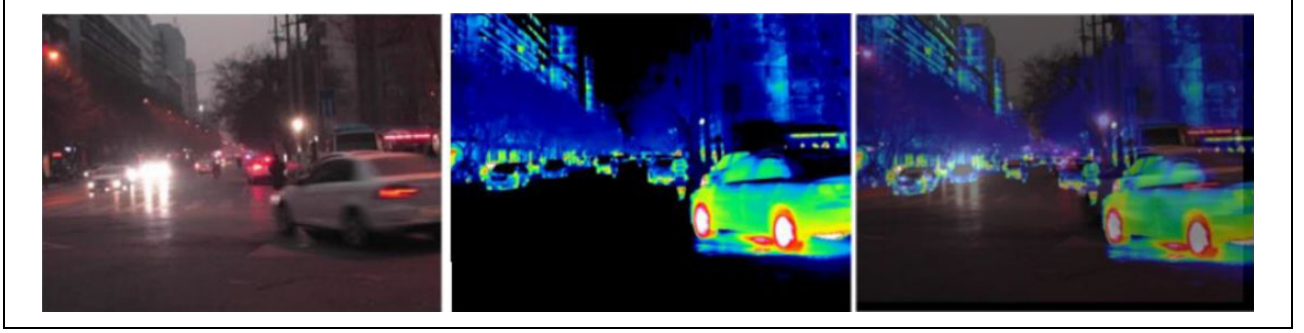


Figure 6. Three-mode images sampled simultaneously.

Table 1. Comparison with traditional SSD series model.

Model	Car _{AP} (%)	Bus _{AP} (%)	mAP (%)
VGG16-SSD	70.7	90.1	80.4
VGG16-FSSD	73.9	89.3	81.6
VGG16-RFBNet	68.3	88.7	78.5
VGG16-Ours	76.9	90.7	83.8

SSD: Single Shot Multi-Box Detector; mAP: mean average precision.

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad i \in N \quad (14)$$

where N is the number of categories.

Before air-to-ground detection, it's necessary to retrain the feature extraction subnetwork to improve the task description performance. When compared to the existed traditional feature description network, such as VGG16, the fine-tune process with the special data sets is able to improve the data adaptability. Therefore, this article utilizes the white light data set to validate the effectiveness of the proposed bidirectional network. And the results are shown in Table 1.

As illustrated in Table 1, we can find that the proposed model, which adopts the bidirectional structure, effectively improves the detection accuracy of air-to-ground small targets. In the following experiments, this article implements the proposed model in three different modal data sets. Herein, the data sets are divided into single-modal data set and multimodal ones. In the single-modal data sets, both the visible data sets and IR data sets are included. Firstly, we use the multimodal data set to improve the performance of single-mode detection. The evaluation of single-modal data sets is divided into visible light validation and IR validation. The contents of the two validation sets are consistent with that of multimodal data sets, which include visible light, IR, and fused data sets. Then the multimodal and single-modal data are respectively performed. The experimental results are shown in Tables 2 and 3.

As described in the above results, we can easily find that the fusion of multimodal features is beneficial to the object description. And the utilization of multimodal data sets can

Table 2. Comparison of single-mode IR image training and three-modal joint training.

Modal type	IR – car _{AP} (%)	IR – bus _{AP} (%)	IR – mAP (%)
IR	79.9	91.5	85.7
Three-modal	80.5	92.1	86.3

mAP: mean average precision.

Table 3. Comparison of single-mode visible light image training and three-modal joint training.

Modal type	Visible light – car _{AP} (%)	Visible light – bus _{AP} (%)	Visible light – mAP (%)
Visible light	76.9	90.7	83.8
Three-modal	78.7	91.1	84.9

mAP: mean average precision.

Table 4. Detection results in weak brightness conditions.

Modal type	Low light – car _{AP} (%)	Low light – bus _{AP} (%)	Low light – mAP (%)
Visible light	70.3	87.4	78.9
IR	72.1	92.5	82.3
Fused modal	78.4	93.1	85.8

mAP: mean average precision.

effectively improve the detection performance compared to the single-modal data set.

In order to show the advantages of the IR and the multi-modal observation, especially when the brightness environment is poor (weak brightness), we implement the detection experiments in the data sets collected in the low-light conditions. The experiment results are shown in Table 4.

From the above results, we can find that the low-light data set verification accuracy for visible light alone is not as high as the picture-in-picture data set. The fused modal data sets that combines visible light and IR can effectively solve the problem of detecting weaker scenes. For the sake of simplicity, the visible results of the detection experiments in weak brightness conditions are shown in Figure 7.

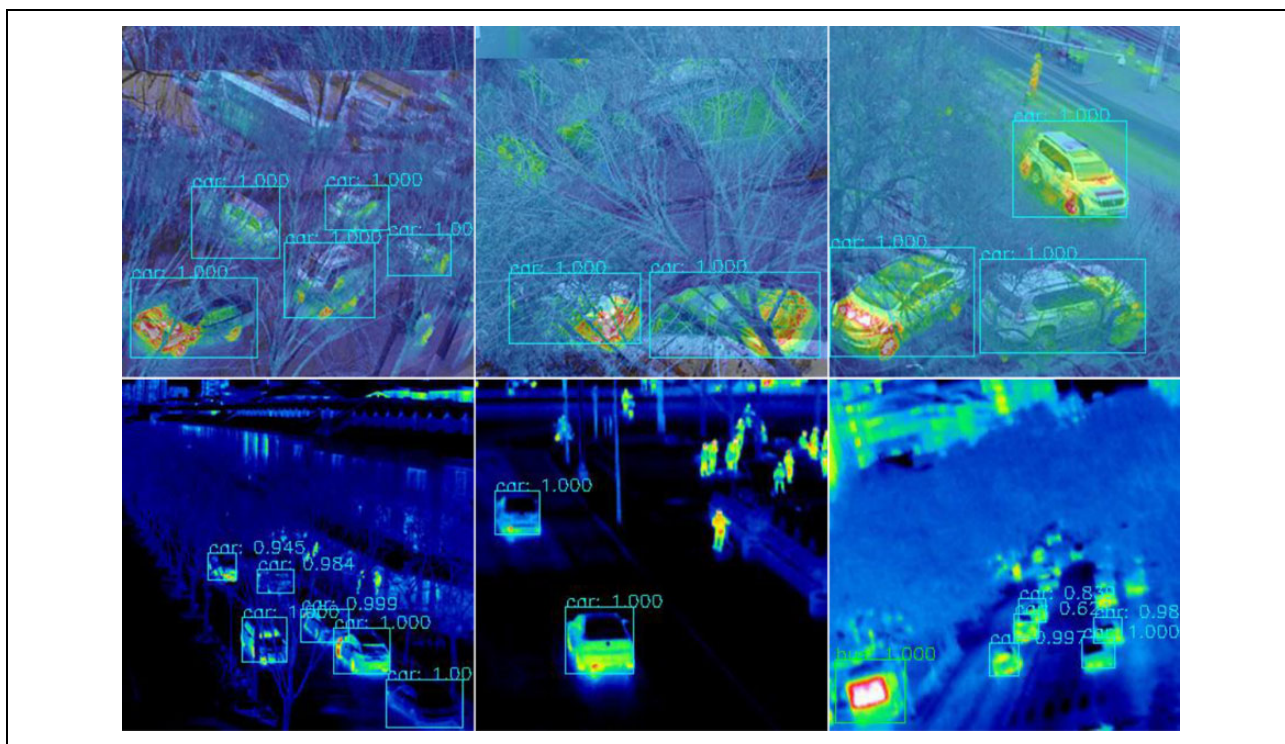


Figure 7. Visualization of the fused modal (top) and IR (bottom) image validation sets.

From the visualization results in Figure 7, it can be seen that the introduction of IR images makes the target features clearer under both low-light conditions and dark conditions. In addition, due to the utilization of infrared sensors, the air-to-ground detection also shows better robustness under the occlusion of trees and other objects.

Conclusions

Air-to-ground object detection are of critical technologies in various application in either military or civilian fields. In aerial observations, the information of target itself is inadequate to interference from large visual complex image. Besides, in case of light deficiency, the performance of object detection based on single-mode visible light is seriously degraded. In order to solve the aforementioned two problems, this article provides improvements from both network structure and data sets. Based on the context, the target detection network with feature association is designed. Infrared imaging is combined with the data sets of two different imaging mechanisms of visible light to produce multimodal data sets. The final experiments validate the effectiveness of the network design and the robustness of the multimodal data set under different brightness conditions, which provides a new effective manner for the aerial detection tasks.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by the National Natural Science Foundation of China (grant nos. 61673017, 61403398) and the Natural Science Foundation of Shaanxi Province (grant nos. 2017JM6016, 2018ZDXM-GY-039).

ORCID iD

Xing Liu  <https://orcid.org/0000-0001-8122-7095>

References

1. Lindeberg T. Scale invariant feature transform. *Scholarpedia* 2012; 2012–2021.
2. Dalal N and Triggs B. Histograms of oriented gradients for human detection. In: *CVPR 2005. IEEE computer society conference on computer vision and pattern recognition, 2005*, San Diego, CA, USA, 21–23 September 2005, pp. 886–893.
3. Bay H, Ess A, Tuytelaars T, et al. Speeded-up robust features. *Comput Vis Image Und* 2008; 110(3): 404–417.
4. Krizhevsky A, Sutskever I, and Hinton GE. ImageNet classification with deep convolutional neural networks. In: *International conference on neural information processing systems*, Lake Tahoe, 3–8 December 2012, pp. 1097–1105.
5. Simonyan K and Zisserman A. Very deep convolutional networks for large-scale image recognition In: *International Conference on Learning Representations 2015*, pp. 1–14.
6. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: *IEEE conference on computer vision and pattern*

- recognition, Boston, Massachusetts, 8–10 June 2015, pp. 1–9. IEEE Computer Society.
7. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: *IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, USA, 27–30 June 2016, pp. 770–778. IEEE Computer Society.
 8. Howard AG, Zhu M, Chen B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications. *CoRR* 2017. abs/1704.04861.
 9. Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In: *IEEE conference on computer vision and pattern recognition*, Columbus, Ohio, 24–27 June 2014, pp. 580–587. IEEE Computer Society.
 10. Girshick R. Fast R-CNN. In: *IEEE international conference on computer vision*, Santiago, Chile, 7–13 December 2015, pp. 1440–1448. IEEE Computer Society.
 11. Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. In: *International conference on neural information processing systems*, Boston, Massachusetts, 8–10 June 2015, pp. 91–99. MIT Press.
 12. Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection. In: *Computer vision and pattern recognition*, Las Vegas, NV, USA, 27–30 June 2016, pp. 779–788. IEEE.
 13. Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector. In: *Computer vision – ECCV 2016*, Amsterdam, 8–16 October 2016, pp. 21–37. Springer International Publishing.
 14. Chen C, Liu MY, Tuzel O, et al. R-CNN for small object detection. In: *Asian conference on computer vision*, Taipei, China, 20–24 November 2016, pp. 214–230.
 15. Liu W, Rabinovich A, and Berg AC. ParseNet: looking wider to see better. *Comput Sci* 2015: 1–8.
 16. Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network. In: *International conference of computer vision*, Venice, Italy, 2017, pp. 6230–6239. IEEE Computer Society.
 17. LI Z and Zhou F. FSSD: feature fusion single shot multibox detector. *CoRR* 2017. abs/1712.00960.
 18. Liu S, Huang D, and Wang Y. Receptive field block net for accurate and fast object detection. In: *European Conference on Computer Vision - ECCV 2018*, Munich, Germany, 2018, pp. 404–419.
 19. Lee K, Choi J, Jeong J, et al. Residual features and unified prediction network for single stage detection. *Comput Sci* 2017.
 20. Fu C Y, Liu W, Ranga A, et al. DSSD: deconvolutional single shot detector. *Comput Sci* 2017.
 21. Kong T, Sun F, Yao A, et al. RON: reverse connection with objectness prior networks for object detection. In: *IEEE conference on computer vision and pattern recognition*, Venice, Italy, 2017, pp. 5244–5252. IEEE Computer Society.
 22. Lin TY, Dollar P, Girshick R, et al. Feature pyramid networks for object detection. In: *IEEE conference on computer vision and pattern recognition*, Venice, Italy, 2017, pp. 936–944. IEEE Computer Society.
 23. Shen Z, Liu Z, Li J, et al. DSOD: learning deeply supervised object detectors from scratch. In: *International conference of computer vision*, Venice, Italy, 2017, pp. 1937–1945.
 24. Ren J, Chen X, Liu J, et al. Accurate single stage detector using recurrent rolling convolution. In: *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, 22–25 July 2017, pp. 752–760.
 25. Jeong J, Park H, and Kwak N. Enhancement of SSD by concatenating feature maps for object detection. *Comput Sci* 2017.
 26. Haber E and Modersitzki J. Intensity gradient based registration and fusion of multi-modal images. *Method Inform Med* 2007; 46(03): 292–299.
 27. Huaping L, Wang F, Sun F, et al. Surface material retrieval using weakly-paired cross-modal learning. *IEEE Trans Autom Sci Eng*, in press. DOI: 10.1109/TASE.2018.2865000.
 28. Niu Y, Lu Z, Wen JR, et al. Multi-modal multi-scale deep learning for large-scale image annotation. *IEEE Trans Image Proc* 2019; 28(4): 1720–1731.
 29. Huaping L, Sun F, and Zhang X. Robotic material perception using active multi-modal fusion. *IEEE Trans Ind Electron*, in press. DOI: 10.1109/TIE.2018.2878157.
 30. Everingham M and Winn J. The PASCAL visual object classes challenge 2010 (VOC2010) development kit contents. In: *International conference on machine learning challenges: evaluating predictive uncertainty visual object classification*, 2011, pp. 117–176. Springer-Verlag.