Multispectral Sample Augmentation and Illumination Guidance for RGB-T Object Detection by MMDetection Framework

Jinqi Yang[1], Xin Yang[1], Yizhao Liao[1], Jinxiang Huang[1], Hongyu He[1], Erfan Zhang[1], Ya Zhou[1] and Yong Song[1,*]

1 School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China; jinqiyang@bit.edu.cn; xinyang@bit.edu.cn; liaoyizhao@bit.edu.cn; huangjinxiang@bit.edu.cn; hehongyu@bit.edu.cn; zhangerfan@bit.edu.cn; zhouya@bit.edu.cn;

* Correspondence: yongsong@bit.edu.cn

Abstract. Multispectral object detection technology has important application prospects in the fields of autonomous driving and so on. Conventional multispectral object detection algorithm rely solely on deep neural networks to learn multispectral image sample information, lacking the guidance of prior knowledge, and not fully utilizing infrared, visible, and other spectral information, resulting in decreased accuracy of object detection in complex scenes. To address this problem, this paper proposes an object detection algorithm based on infrared visible sample augmentation and illumination guidance. The algorithm adopts the MMDetection framework and extracts multispectral object features based on a designed sample augmentation method based on the fusion of positive and negative samples in multispectral images. Based on a designed adaptive weight allocation method guided by illumination, it enhances the algorithm's adaptability to the lighting environment. Finally, through the design of a multi-task loss function, it achieves high-precision and robust object detection in complex scenes. Experimental results on datasets such as FLIR and M^{3}FD show that the proposed algorithm has significant advantages over comparative algorithms such as CFR_3 and GAFF in terms of average detection precision.

Keywords: machine vision; object detection; feature fusion; multispectral

1. Introduction

As a key task in the field of computer vision applications, object detection[1] has broad application prospects, such as safety monitoring, intelligent driving, unmanned aerial vehicles, etc. Traditional multispectral object detection algorithms require manually designed features, such as using detection algorithms based on multispectral aggregation features ACF+T+THOG[2]. Although easy to deploy, it is difficult to extract discriminative features and cannot meet high-precision requirements. The emergence of deep learning based multispectral object detection algorithms has largely solved the above problems. Two-stage detection methods such as Liu[3] designed four different fusion architectures and proved that the performance obtained by mid-term fusion is the best. Although the above two-stage methods have higher accuracy, the one-stage method has a higher advantage in terms of speed. Zhang H[4] proposed CFR_3 algorithm based on cyclic fusion to enhance feature consistency between different modalities, effectively achieving consistency/complementarity balance in multispectral features. Zhang

This work is supported by National Natural Science Foundation of China General Program (82272130) National Natural Science Foundation of China Key Program (U22A20103) and Aeronautical Science Foundation (2023Z019072001).

H[5] proposed GAFF, combining intra and inter modal attention to learn multimodal features. Current methods rely on DNNs, ignoring prior knowledge and underutilizing infrared and visible information, leading to reduced accuracy in complex scenes. The main ideas of the algorithm in this paper include:

1. This paper proposes a multispectral object detection algorithm that combines infrared visible sample augmentation and illumination guidance technology.
2. We design a sample augmentation method based on the fusion of multispectral positive and negative samples, extracting features of multispectral targets such as infrared and visible. According to the designed positive and negative sample fusion strategy, the training dataset of infrared visible images is effectively expanded.
3. We design an adaptive weight allocation module based on illumination guidance, which dynamically adjusts the weights of infrared-visible modal features according to changes in scene illumination.

2. Method

As shown in figure 1, the proposed network structure of the multispectral object detection algorithm based on RGB-T data augmentation and illuminance guidance, which primarily consists of the following components: a multispectral data augmentation module (MDA) for positive and negative sample fusion, an illuminance guidance module (IG), a feature extraction and feature fusion network, and the detector.

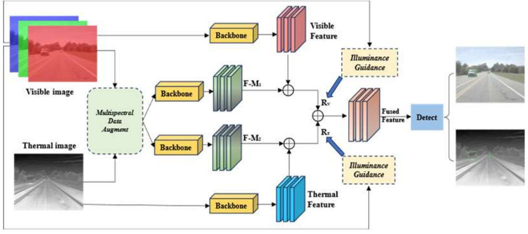

Figure.1 Object detection network based on infrared-visible sample augmentation and illuminance guidance.

2.1. Data augmentation based on multispectral positive and negative sample fusion

Firstly, a pair of infrared-visible image pairs are randomly selected with a width and height of *W* and *H*, respectively. The size of the concatenated images is kept consistent with the original image. The specific augmentation methods are as follows: for figure2 (a)(b)(c), a clipping coefficient *k* is defined, whose value is taken from [0.3,0.7], while for figure2 (d), the number of clipping is defined as *n*, where *n* ≤ the number of objects:

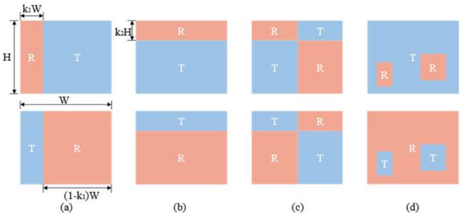

Figure.2 Infrared-visible sample augmentation. (a) horizontal augment (b) vertical augment (c) diagonal augment (d) foreground/background augment.

Vertical and horizontal augmentation involve cropping infrared and visible image pairs, based on scaling coefficient *k*, and concatenating with subblocks of different modalities. Diagonal augmentation crops image pairs into four subblocks with random ratios and arranges modal subblocks adjacently, while same modal subblocks diagonally. Cutting line control is necessary to prevent target cutting. Foreground/background augmentation extracts targets from infrared-visible image pairs, replaces them with images of different modalities, maintaining relative positions. This approach leverages clearer backgrounds from visible images and infrared sensitivity to objects, enhancing both background and target information. It improves detection accuracy and robustness in complex backgrounds by learning modality consistency and differences.

2.2. Adaptive weight allocation module based on illumination guidance

The TINet[6] lighting subnet predicts daytime or nighttime scenes in visible images and weights modal features, but its label division has problems and does not consider how to allocate labels in the evening and rainy days. Due to the majority of scenes are traffic in the center of the road, the object is concentrated in the middle area of the image, while the sky and ground colors in the upper and lower parts of the image are relatively single. It is often easy to determine the environment of the image by judging the peak values of the grayscale histogram in this part.

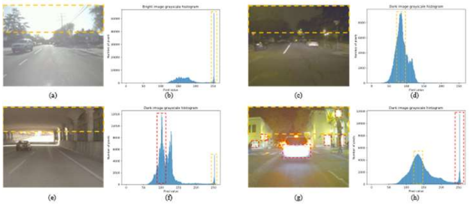

Figure.3 Images of different illuminance and grayscale histograms. (a)(c)(e)(g) are visible images with different illuminances, while (b)(d)(f)(h) are the corresponding grayscale histogram.

As shown in Figure 3, it is easy to judge the lighting situation of bright daytime images and dim nighttime images based on the grayscale histogram of the entire image, as shown in Figures3 (b) and (d). However, mis-judgment is prone to occur during the day in dim tunnels or in scenes with light pollution at night, as shown in Figures3 (f) and (g).

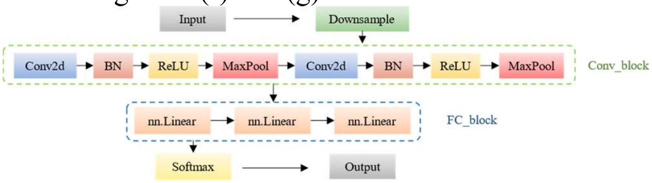

Figure.4 The network structure diagram of adaptive weight distribution module guided by illuminance

After determining the lighting label, the illumination guidance module is used to obtain the probability of illumination prediction and calculate the loss. The network structure of the illumination guidance module is shown in Figure 4. The softmax operation is performed to obtain the predicted probability *P_D* and *P_N* of the image scene being dark or bright. In order to better balance the contribution of the two modalities, the weight is calculated according to equation (2.1):

$$\begin{cases} R_V = \frac{1}{2} \cdot \left[(P_D - P_N) \cdot \sigma + 1 \right] \\ R_T = 1 - R_V \end{cases} \tag{2.1}$$

R_V and R_T represent the weights of visible and infrared images, σ is learnable parameter.

2.3. Multi-task loss function

The multi-task loss L includes the MDA loss function L_{mda}, the IG loss function L_g, the classification loss function L_{cls}, the localization loss function L_{reg}, and the penalty coefficient α. L is given by

$$L = \alpha_1 L_{\mathrm{mda}} + \alpha_2 L_{\mathrm{g}} + \alpha_3 L_{cls} + \alpha_4 L_{reg} \tag{2.2}$$

The multispectral data augmentation loss function L_{mda} use Smooth L1 loss:

$$L_{\mathrm{mda}}(x) = \begin{cases} 0.5x^2, & |x| < 1 \\ |x| - 0.5, & x < -1 \text{ or } x > 1 \end{cases} \tag{2.3}$$

Where x is the proportion of data augmentation and cropping.

The loss function of the illumination guidance module is cross entropy loss:

$$L_{\mathrm{g}} = -\frac{1}{N} \sum_{i=1}^{N} (y_i \log(p_i) + (1 - y_i) \log(1 - p_i)) \tag{2.4}$$

Where N represents the number of samples in one batch, y_i is the classification label of the ith input image, and p_i is the probability that the ith input image is a daytime image.

The classification and regression loss functions are the same as the Faster R-CNN[7].

3. Experiment

Experimental Setup. All model training is conducted on the RTX 2080TI, and the proposed algorithm is implemented based on Pytorch[8] and MMDetection[9]. The model training is conducted using non-distributed training and evalution use a single GPU. The code is based on MMDetection2.4.0 and Pytorch 1.7.1 in Python 3.7.

Dataset. FLIR-aligned[10]: an aligned version based on the FLIR dataset. There are a total of 5142 pairs of image pairs. The scene includes vehicles and pedestrians on the road with a size of 640×512. Including three categories: cars, people, and bicycles; M^3FD[11]: currently the latest multispectral object detection dataset. It consists of 4200 aligned image pairs. An infrared-visible image pair with a size of 1024×768 is obtained. Including six categories: people, car, bus, motorcycle, truck, lamp.

Training Settings. We use Faster R-CNN as the basic detector. During training, the optimizer uses the SGD, setting the initial learning rate to 5×10^{-3} . Batchsize set to 4, and a total of 12 epochs are trained. The probability of vertical, horizontal, diagonal and fore/background augmentation is 0.15, 0.1, 0.05 and 0.05. During training, the network backbone uses pre-trained ResNet50 [12] and FPN [13].

3.1. Ablation Study

From Table 1, it can be seen that on the FLIR-aligned dataset, compared to the initial M1 model, the M2 model with only IG increased its AP_{50} by 1.1%. The AP_{50} of the M3 model increased by 1.4% after adding MDA. The AP_{50} of the M5 model designed in this paper has increased by 2.5%.

On the M^3FD dataset, compared to the initial M1 model, when only IG is used, the AP_{50} increases by 0.3%. After adding MDA, the AP_{50} relatively increased by 1.4%. When three improvements are adopted simultaneously, the AP_{50} increases by 2.8%. This indicates that the designed LG, MDA, and M-Loss methods have effectively improved the accuracy of the multispectral object detection network.

Table 1. Ablation experiment results on FLIR-aligned and M^3FD datasets.

Model	Module			Dataset			
				FLIR-aligned		M^3FD	
	LG	MDA	M-Loss	mAP	AP_{50}	mAP	AP_{50}
M1				37%	76%	41.1%	64.6%
M2	√			37.8%	76.8%	41.6%	64.8%
M3		√		38.2%	77.1%	42%	65.5%
M4	√	√		38.5%	77.3%	42.5%	65.7%
M5	√	√	√	39.1%	77.9%	43.1%	66.4%

3.2. Quantitative Analysis

We compared our algorithm with state-of-the-art object detectors such as Faster R-CNN, SSD, HalfwayFusion, CFR3, GAFF, ThermalDet, YOLOv5s, YOLOF, TINet, etc.

Table 2 shows the experimental results of our algorithm and the state-of-the-art algorithm on this dataset. Our proposed algorithm achieves an AP_{50} of 77.9%, which is not only superior to other advanced detection algorithms compared, but also an AP_{50} improvement of 2.4% compared to suboptimal algorithms. It is worth mentioning that our algorithm achieved the best detection results in both the key target categories of car and person.

According to Table 3, the algorithm proposed in this paper achieved the best precision on the M^3FD dataset, except for the Lamp category. Compared to the basic detector Faster R-CNN, the AP_{50} increased from 62.03% to 66.4%.

Table 2. Comparison of detection results (AP_{50}) on FLIR-aligned datasets.

Method	Car	Person	Bicycle	AP_{50}↑
Faster R-CNN[7]	67.60	39.60	54.70	53.90
SDD [14]	61.60	40.90	43.60	48.70
HalfwayFusion [3]	—	—	—	71.20

CFR_3 [4]	84.91	74.49	55.77	72.39
GAFF [5]	—	—	—	72.90
ThermalDet [15]	85.52	*78.24*	60.04	74.60
YOLOv5s [16]	80.00	68.30	*67.10*	71.80
YOLOF [17]	79.40	67.80	***68.10***	71.81
TINet[6]	*86.04*	75.57	54.63	*76.07*
Ours	**88.36**	**81.32**	64.08	**77.90**

Table 3. Comparison of detection results (AP_{50}) on M^3FD datasets.

Method	People	Car	Bus	Motorcycle	Truck	Lamp	$AP_{50}\uparrow$
Faster R-CNN [7]	59.86	*81.52*	*79.96*	*47.94*	66.50	36.14	62.03
YOLOv3 [18]	40.52	61.82	61.20	31.85	30.16	25.94	41.90
YOLOF [17]	27.72	53.18	56.53	13.16	38.21	11.57	33.40
Retinanet[19]	53.57	72.04	65.91	20.1	46.4	21.76	46.63
FCOS[20]	*61.32*	75.13	79.69	41.23	*74.93*	**45.16**	*62.91*
Ours	**62.50**	**83.03**	**84.41**	**54.59**	**75.52**	*37.73*	**66.40**

(In Table 2 and Table 3, the values highlighted in red and bold are the best results, the values highlighted in blue and italics are the suboptimal results, and the underlined values are the third-ranked results)

3.3. Qualitative Analysis

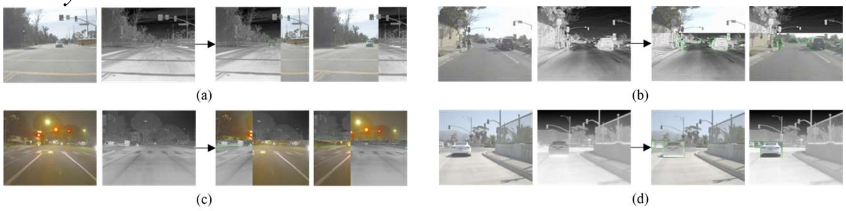

Figure 5. Instance graphs of multispectral data augmentations on FLIR-aligned datasets. (a) Horizontal augment (b) Vertical augment (c) Diagonal augment (d) Fore/Background augment.

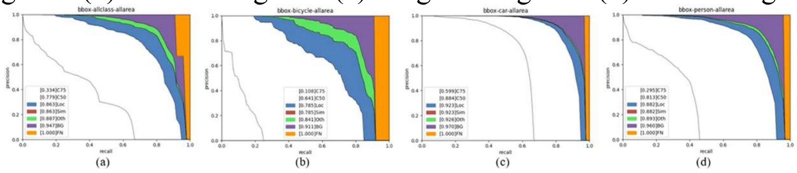

Figure 6. PR curves for all target categories on the FLIR-aligned dataset.

Figure 5. demonstrate the effectiveness of the data augmentations designed in this paper more clearly, some image pairs in the FLIR-aligned dataset were visualized. It shows the infrared visible image pairs and their detection results after several data augmentations under different lighting conditions. It can be seen that the object cars and people in the figure have been mostly detected.

As shown in Figure 6, C75 and C50 are the PR values at IoU=0.75 and 0.5, respectively. Loc is the PR value at IoU=0.1. The FN value is the PR value after removing all remaining errors (AP=1). Figure 5 (a) shows that among all categories, localization accuracy has the greatest impact, followed by confusion with the background. Comparing Figure 5 (b), (c), and (d), it can be seen that the detection performance of car is the best, while the detection performance of bicycle is relatively poor.

4. Conclusion

This paper proposes an object detection algorithm based on infrared visible sample augmentation and illumination guidance. The algorithm adopts the MMDetection framework and extracts multispectral object features based on a designed sample augmentation method based on the fusion of positive and negative samples in multispectral images. Based on a designed adaptive weight allocation method guided by illumination, it enhances the algorithm's adaptability to the lighting environment. Finally, the algorithm achieved an average accuracy of 77.9% on the typical intelligent driving dataset FLIR aligned, and 66.4% on the challenging multispectral object detection dataset M3FD. In addition, the experimental results also proved that each module designed is effective for detection accuracy.

References

[1] Zhao Z Q, Zheng P, Xu S T, et al. (2019) Object detection with deep learning: a review. *TNNLS*, vol.30, pp.3212–3232.

[2] Hwang S, Park J, Kim N, et al. (2015) Multispectral pedestrian detection: Benchmark dataset and baseline. *CVPR*, pp. 1037–1045. DOI: 10.1109.

[3] Liu J, Zhang S, Wang S, Metaxas D N. (2016) Multispectral deep neural networks for pedestrian detection. arXiv:1611.02644.

[4] Zhang H, Fromont E, Sébastien Lefevre, et al. (2020) Multispectral fusion for object detection with cyclic fuse-and-refine blocks. *ICIP*, pp. 276-280. DOI:10.1109/ICIP40778.2020.9191080.

[5] Zhang H, Fromont E, Lefevre S, et al. (2021) Guided Attentive Feature Fusion for Multispectral Pedestrian Detection. *WACV*, pp. 72-80. DOI:10.1109.

[6] Zhang Y, Yu H, He Y, et al. (2023) Illumination-guided RGBT object detection with inter- and intra-modality fusion. *TIM*, vol.72, pp.1-13.

[7] Ren S, He K, Girshick R, et al. (2017) Faster R-CNN: Towards real-time object detection with region proposal networks. *TPAMI*, pp. 1137-1149. DOI: 10.1109.

[8] Paszke A, et al. (2019) PyTorch: An imperative style, high-performance deep learning library. *NIPS*, pp.8026–8037.

[9] Chen K, Wang J, Pang J, et al. (2019) MMDetection: Open MMLab detection toolbox and benchmark. arXiv:1906.07155.

[10] FLIR, 2022. Free Teledyne FLIR thermal dataset for algorithm training. [Online]. Available: https://www.flir.com/oem/adas/adas-dataset-form/.

[11] Liu J, Fan X, Huang Z, et al. (2022c) Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. *CVPR*, pp.5802–5811.

[12] He K, Zhang X, Ren S, et al. (2016) Deep residual learning for image recognition. *CVPR*, pp. 770–778.

[13] Lin T Y, Dollar P, Girshick R, et al. (2017) Feature pyramid networks for object detection. *CVPR*, pp.2117–2125.

[14] Berg A C, Fu C Y, Szegedy C, et al. (2015) SSD: Single Shot MultiBox Detector. *ECCV*, vol.9905. DOI:10.1007.

[15] Cao Y, Zhou T, Zhu X, Su Y. (2019) Every Feature Counts: An Improved One-Stage Detector in Thermal Imagery. *ICCC*, pp.1965–1969.

[16] Glenn J, Ayush C, Alex S, et al. (2022) ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation. *Zenodo*.

[17] Chen Q, Wang Y, Yang T, et al. (2021) You Only Look One-level Feature. *CVPR*, pp.13034–13043.

[18] Redmon J, Farhadi A. (2018) YOLOv3: An Incremental Improvement. arXiv:1804.02767. DOI:10.48550.

[19] Lin T Y, Goyal P, Girshick R, et al. (2017) Focal Loss for Dense Object Detection. *TPAMI*, pp.2999-3007.

[20] Tian Z, Shen C, Chen H, et al. (2020) FCOS: Fully Convolutional One-Stage Object Detection. *ICCV*, pp.9626-9635.

[21] Song K, Zhao Y, Huang L, et al. (2023) RGB-T image analysis technology and application: A survey. *EAAI*, vol.120, pp.105919.

[22] Wang Q, Chi Y, Shen T, Song J, Zhang Z, Zhu Y. (2022) Improving RGB-Infrared Object Detection by Reducing Cross-Modality Redundancy. *RS*, pp.2020.

[23] Chen K, Liu J, Zhang H. (2023) IGT: Illumination-guided RGB-T object detection with transformers. *KBS*, pp.110423.

[24] Liu Y, Zeng Y, Qin J. (2024) GSC-YOLO: a lightweight network for cup and piston head detection. *SIViP*, pp. 351–360.

[25] Hou Z, Yang C, Sun Y, Ma S, Yang X, Fan J. (2024) An object detection algorithm based on infrared-visible dual modal feature fusion. *IPT*, pp.105107.