ICIC Express Letters
Volume **18**, Number **3**, March **2024**

 ISSN 1881-803X
pp. **241–245**

YOLOV7-BASED OBJECT DETECTION USING RGB AND IR IMAGES

Kosuke Shigematsu* and Rentaro Kato

Department of Information Engineering
National Institute of Technology, Oita College
1666 Maki, Oita City, Oita 870-0152, Japan
reind0706@gmail.com
*Corresponding author: k-shigematsu@oita-ct.ac.jp

Received October 2023; accepted December 2023

Abstract. *Object detection based on deep learning typically relies on visible images. However, environmental factors such as nighttime, can significantly degrade performance. Multimodal object-detection techniques that combine visible and infrared (IR) images can be effective in addressing this issue. However, multimodal networks generally incur higher computational costs than processing visible images alone. In this study, we propose a network that builds upon YOLOv7, which is renowned for its high performance in real-time object detection, to reduce computational load while improving performance. Specifically, the initial convolutional layers of YOLOv7 are fine-tuned to handle 4-channel data, comprising RGB and IR. This approach offers the advantage of concurrently minimizing the number of network parameters and enhancing the accuracy of object detection. The experimental results obtained on two datasets, low-light visible-infrared paired (LLVIP) and Multispectral datasets, demonstrate the superiority of this new method.*
Keywords: Object detection, Infrared image, Visible image, YOLOv7

1. **Introduction.** In recent years, object detection technologies employing deep learning have been utilized across a diverse range of fields, such as autonomous driving and surveillance cameras, and infrastructure inspection [1]. Among these, visible cameras are widely used; however, their performance is significantly affected by environmental conditions such as rain, fog, and low-light that substantially impair object detection accuracy.

To solve this problem, ensemble methods have been employed to fuse the inference results of object detection from visible images with those of infrared images [2]. Recently, multimodal object-detection methods that use both visible and far-infrared images have also been proposed [3, 4]. These methods reportedly demonstrate higher performance than using visible images alone, improving object detection accuracy in outdoor environments. However, these approaches generally entail higher computational costs than those using only visible images, which can be a significant issue, particularly in applications requiring real-time processing.

The YOLO series of object detection networks are widely used due to their real-time processing capabilities and excellent performance. In their research, Tu et al. have enhanced the object detection performance for both visible and infrared images by applying contrastive learning to YOLOv5 [5].

Among multimodal networks, a method has been proposed, whereby visible and infrared images are integrated prior to inputting them into the network [6]. This method is based on the YOLOv3 object-detection network [7] and mitigates the increase in network size by using an integrated 4-channel image consisting of visible and infrared images as input. In this study, we extend this method and propose a network adapted to YOLOv7

DOI: 10.24507/icicel.18.03.241

[8] that exhibits even higher performance in real-time object detection. The proposed network can handle 4-channel data, including not only visible but also far-infrared images. This approach enables the utilization of features from both visible and infrared images, resulting in enhanced performance compared to using either type of image alone. Furthermore, it efficiently minimizes the increase in computational cost. The experimental results utilizing the LLVIP dataset [9] and the Multispectral dataset [2], both comprising visible and IR images, demonstrated superior performance over existing methods across several performance metrics.

2. **Proposed Method.** In this paper, we propose a network architecture based on YOLOv7 with 4-channel (RGB and IR) image input. This network handles images with four channels, incorporating infrared in addition to the standard Red, Green, Blue channels. By extending the input image to four channels, more abundant information than RGB or IR images alone can be fed as input, to improve performance. In addition, building upon the network structure of YOLOv7, makes real-time processing feasible. An overview of the proposed method is presented in Figure 1.

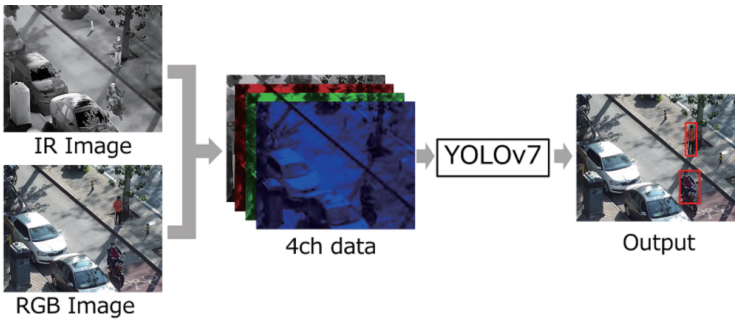

FIGURE 1. Overview of the proposed method

In the proposed method, minor adjustments were made to the initial convolutional layers of YOLOv7 to add an IR channel. Specifically, 32 3×3 kernels were added to the IR channel, the same number as that in each RGB channel. Consequently, the total number of parameters increased by a mere 288 ($32 \times 3 \times 3$), from 36,481,772 to 36,482,060. This increase rate of approximately 0.00079%, ensures that the increase in computational cost is minimal. We anticipated that this slight parameter adjustment would allow us to capitalize on the benefits of incorporating IR image information, thereby achieving high-precision real-time object detection with minimal computational overhead.

3. **Dataset.**

3.1. **LLVIP dataset.** The low-light visible-infrared paired (LLVIP) dataset [9] consists of paired RGB and infrared (wavelength: 8-14 μm) images, specifically targeting detection tasks under low-light conditions as well as image-to-image translation. The RGB and infrared images were aligned in terms of their positions. The dataset included 15,488 pairs of images with a resolution of 1280×1024 pixels captured during both daytime and nighttime. Of these, 12,025 were used as training data, and 3,463 were allocated as test data. The dataset focused on the single class of 'pedestrian'. An example of the LLVIP dataset is shown in Figure 2.

3.2. **Multispectral dataset.** The Multispectral dataset [2] includes RGB, far-infrared (FIR), mid-infrared (MIR), and near-infrared (NIR) images, each positionally aligned. The dataset contains five classes: 'bike', 'car', 'car stop', 'color cone', and 'person'. Of the 2,999 images included in the dataset, 500 were designated for testing and 2,499 for training. Each image had a resolution of 320×256 pixels. An example from the Multispectral dataset is shown in Figure 3.

(a) RGB image (b) IR image

FIGURE 2. An example from the LLVIP dataset [9]

(a) RGB image (b) NIR image

(c) MIR image (d) FIR image

FIGURE 3. Example from the Multispectral dataset [2]

4. **Result.**

4.1. **LLVIP dataset.** The comparative results of the proposed method with those of previous studies are discussed on the LLVIP dataset. The evaluation results are presented in Table 1 for three different input datasets: IR alone, RGB alone, and 4-channel (RGB and IR). The training and inference were performed at a resolution of 1280×1280. Default

TABLE 1. Performance evaluation on the LLVIP dataset using various YOLOv7 configurations

Model	AP50 (%)	AP75 (%)	AP50:95 (%)	Parameters
CFT [3]	97.5	72.9	63.6	–
Infusion-Net [4]	**98.6**	73.3	64.6	–
YOLOv7 (IR)	97.9	78.2	67.2	36,481,772
YOLOv7 (RGB)	92.6	58.8	54.4	36,481,772
YOLOv7 (4ch: RGB and IR)	97.9	**78.3**	**67.2**	36,482,060

parameter values were used for data augmentation. Model performance was evaluated using three metrics: average precision at 50% IoU (AP50), average precision at 75% IoU (AP75), and mean average precision (mAP) calculated with an intersection over union (IoU) threshold that varied from 50% to 95% in increments of 5% (AP50:95). AP50 is the average precision where a bounding box proposed by an object detection model is deemed correct if its Intersection over Union (IoU) with the actual bounding box exceeds 50%. Similarly, AP75 uses a higher IoU threshold of 75%, evaluating the model's accuracy with more rigorous criteria. AP50:95 calculates the average AP at each IoU threshold, incrementing by 5% from 50% to 95%, thus gauging the model's comprehensive performance. These metrics serve as indicators of object detection accuracy, with higher values indicating better performance.

According to the evaluation results, a slight improvement was observed in AP75 when using the 4-channel (RGB and IR) data compared to IR alone. Furthermore, the 4-channel (RGB and IR) approach outperformed previous studies for both the AP75 and AP50:95 metrics. From these results, it is evident that the proposed YOLOv7 model using 4-channel data (RGB and IR) successfully integrates information from both RGB and IR sources to achieve high-precision object detection, outperforming existing methods in both AP75 and AP50:95 metrics. Regarding the number of parameters, the introduction of the 4th channel resulted in only a slight increase with real-time performance maintained.

4.2. **Multispectral dataset.** Comparative results are presented for the proposed method and those of existing studies on the Multispectral dataset, as detailed in Table 2. As in the case of the LLVIP dataset, we evaluated the performance using mAP50, mAP75, and mAP50:95 as metrics, performing training and inference for different input scenarios: RGB alone, NIR alone, MIR alone, FIR alone, 4-channel (RGB and NIR), 4-channel (RGB and MIR), and 4-channel (RGB and FIR). The image size used for both training and inference was 640×640. Default parameter values were used for data augmentation.

TABLE 2. Performance evaluation on the Multispectral dataset

Model	mAP50 (%)	mAP75 (%)	mAP50:95 (%)	Parameters
YOLOv7 (RGB)	58.6	15.5	22.0	36,481,772
YOLOv7 (NIR)	52.1	14.7	22.6	36,481,772
YOLOv7 (MIR)	56.7	20.0	25.2	36,481,772
YOLOv7 (FIR)	48.4	16.5	22.7	36,481,772
YOLOv7 (4ch: RGB and NIR)	56.2	26.9	28.9	36,482,060
YOLOv7 (4ch: RGB and MIR)	57.3	**27.9**	**29.8**	36,482,060
YOLOv7 (4ch: RGB and FIR)	**59.8**	23.0	29.7	36,482,060

Performance improvements were observed in the AP75 and AP50:95 metrics when using 4-channel (RGB and NIR) and 4-channel (RGB and MIR) images compared to using RGB, NIR, MIR, and FIR images alone. Furthermore, the 4-channel (RGB and FIR) configuration outperformed the individual performances of RGB, NIR, MIR, and FIR images

across all metrics. These results indicate that regardless of the IR wavelength, integrating RGB and IR images leads to performance improvements, confirming the effectiveness of the proposed method.

5. **Conclusions.** In this study, we proposed and evaluated a new network based on YOLOv7 that uses both RGB and infrared (IR) images as inputs to address the problem of reduced object detection accuracy caused by environmental factors such as lighting conditions. This network can integrate 4-channel (4ch) images by combining RGB and IR and provides richer information than using either RGB or IR images alone. The method enables high object detection accuracy while maintaining real-time processing capabilities with minimal increase in the number of parameters.

Experiments were conducted on the LLVIP and Multispectral datasets to confirm the effectiveness and superiority of the proposed method over existing approaches. On both datasets, the method demonstrated higher performance than when using either RGB or IR images alone. Additionally, on the LLVIP dataset, the proposed method outperformed existing methods on the AP75 and AP50:95 evaluation metrics.

Based on these results, it can be concluded that the proposed method is useful for a wide range of applications requiring real-time processing, such as autonomous vehicles and surveillance systems. Although this study integrated RGB and IR for the 4-channel input, the performance can be potentially improved by increasing the number of input channels. Moreover, the network used in this study is not only applicable to pairs of visible and infrared images but can also be adapted to different types of image pairs, such as visible and depth images. Future research is anticipated to conduct experiments to assess the effectiveness and performance of the network with various image pairings. It is expected that adopting newer real-time object detection architectures [10] as the base, will provide further improvements in accuracy and speed.

REFERENCES

[1] L. Guo, R. Li and B. Jiang, A road surface damage detection method using YOLOv4 with PID optimizer, *International Journal of Innovative Computing, Information and Control*, vol.17, no.5, pp.1763-1774, 2021.

[2] T. Karasawa, K. Watanabe, Q. Ha, A. Tejero-De-Pablos, Y. Ushiku and T. Harada, Multispectral object detection for autonomous vehicles, *The 25th Annual ACM International Conference on Multimedia (ACMMM 2017)*, 2017.

[3] Q. Fang, D. Han and Z. Wang, Cross-modality fusion transformer for multispectral object detection, *arXiv Preprint*, arXiv: 2111.00273, 2021.

[4] J.-S. Yun, S.-H. Park and S. B. Yoo, Infusion-Net: Inter- and intra-weighted cross-fusion network for multispectral object detection, *Mathematics*, vol.10, no.21, 3966, DOI: 10.3390/math10213966, 2022.

[5] X. Tu, Z. Yuan, B. Liu, J. Liu, Y. Hu, H. Hua and L. Wei, An improved YOLOv5 for object detection in visible and thermal infrared images based on contrastive learning, *Frontiers in Physics*, vol.11, 2023.

[6] V. Knyaz, Multimodal data fusion for object recognition, *Proc. of SPIE 11059, Multimodal Sensing: Technologies and Applications*, vol.11059, DOI: 10.1117/12.2526067, 2019.

[7] J. Redmon and A. Farhadi, YOLOv3: An incremental improvement, *arXiv Preprint*, arXiv: 1804.02767, 2018.

[8] C.-Y. Wang, A. Bochkovskiy and H. Y. M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.7464-7475, 2023.

[9] X. Jia, C. Zhu, M. Li, W. Tang and W. Zhou, LLVIP: A visible-infrared paired dataset for low-light vision, *Proc. of the IEEE/CVF International Conference on Computer Vision*, pp.3496-3504, 2021.

[10] W. Lv, Y. Zhao, S. Xu, J. Wei, G. Wang, C. Cui, Y. Du, Q. Dang and Y. Liu, DETRs beat YOLOs on real-time object detection, *arXiv Preprint*, arXiv: 2304.08069, 2023.