Hindawi
Wireless Communications and Mobile Computing
Volume 2023, Article ID 9830151, 1 page
https://doi.org/10.1155/2023/9830151

Retraction

Retracted: Research on Multimodal Image Fusion Target Detection Algorithm Based on Generative Adversarial Network

Wireless Communications and Mobile Computing

Received 17 October 2023; Accepted 17 October 2023; Published 18 October 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] Z. Wu, X. Wu, Y. Zhu et al., "Research on Multimodal Image Fusion Target Detection Algorithm Based on Generative Adversarial Network," *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 1740909, 10 pages, 2022.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 1740909, 10 pages
https://doi.org/10.1155/2022/1740909

Research Article

Research on Multimodal Image Fusion Target Detection Algorithm Based on Generative Adversarial Network

Zhaoli Wu,[1,2,3,4] **Xuehan Wu**,[2,3] **Yuancai Zhu**,[2,3,4] **Jingxuan Zhai**,[2,3,4] **Haibo Yang**,[2] **Zhiwei Yang**,[2] **Chao Wang**,[2] **and Jilong Sun**[2]

[1]*China University of Mining and Technology, School of Computer Science and Technology, Xuzhou 221116, China*
[2]*Jiangsu Vocational Institute of Architectural Technology, School of Information and Electronics Engineering, Xuzhou 221116, China*
[3]*Xuzhou Intelligent Machine and Visual Application Technology Engineering Research Center, Xuzhou 221116, China*
[4]*Xuzhou Big Data Analysis and Data Security Engineering Research Center, Xuzhou 221116, China*

Correspondence should be addressed to Zhaoli Wu; lb20170009@cumt.edu.cn

Received 7 December 2021; Revised 21 December 2021; Accepted 27 December 2021; Published 24 January 2022

Academic Editor: Liqin Shi

In this paper, we propose a target detection algorithm based on adversarial discriminative domain adaptation for infrared and visible image fusion using unsupervised learning methods to reduce the differences between multimodal image information. Firstly, this paper improves the fusion model based on generative adversarial network and uses the fusion algorithm based on the dual discriminator generative adversarial network to generate high-quality IR-visible fused images and then blends the IR and visible images into a ternary dataset and combines the triple angular loss function to do migration learning. Finally, the fused images are used as the input images of faster RCNN object detection algorithm for detection, and a new nonmaximum suppression algorithm is used to improve the faster RCNN target detection algorithm, which further improves the target detection accuracy. Experiments prove that the method can achieve mutual complementation of multimodal feature information and make up for the lack of information in single-modal scenes, and the algorithm achieves good detection results for information from both modalities (infrared and visible light).

1. Introduction

With the rapid development of deep learning, the task of target detection in computer vision tasks has made great progress. However, the task of target detection is very difficult to apply in some real-world scenarios. In the military, security, and other fields, traditional visible light images have very obvious limitations. In recent years, scholars have found that the introduction of multimodal data can significantly improve the accuracy of detection algorithms. Multimodality refers to image pairs formed by applying different imaging principles to the same scene. With the successful application of deep convolutional neural networks in target detection tasks, scholars have produced many excellent results in multimodal research. The author uses a convolutional neural network to fuse two modal information and discusses the impact of different fusion stages on the target detection results [1]. The author believes that only fusing two modal information for target detection is imperfect, and it is necessary to retain the unique information of the two modalities [2]. Therefore, the author adds two modules to the network based on the idea of probability, and one module is used. To output the degree of dependence of the current image on the respective features and fusion features of the two modalities, the second module uses the output of the module 1 as the weight, and the respective output results of the two modalities and the output results of the fusion feature are weighted to obtain the discrimination probability. Konig et al. use the faster RCNN target detection algorithm, but use the fused feature layer and the two-modal feature layer in the training process. Literature [3] adjusts the fusion weight of each modal under

different lighting conditions by designing a light perception network to simulate day and night illumination, but the detection accuracy is very dependent on the light perception network.

In response to the above problems, this paper starts from the perspective of the adversarial discriminant domain [4], uses an unsupervised learning method to reduce the modal difference between bimodal images, and proposes a modal information fusion detection algorithm based on a generative adversarial network. In the improved generative confrontation network, the generator is designed with local detail features and global semantic features to extract source image details and semantic information, and perceptual loss is added to the discriminator to keep the data distribution of the fused image consistent with the source image and improve fusion image accuracy. The fused features enter the interest pooling network for rough classification, and the generated candidate frame is mapped to the feature map, and finally, the target classification and positioning are completed through the fully connected layer.

2. Algorithm Structure

In traditional infrared and visible light image fusion methods, a hybrid model is usually established to combine the advantages (salience) of multiple parties. Although the image fusion performance is improved, the fusion rules need to be manually designed. Generative adversarial networks (GAN) have inherent advantages in the field of image generation and can fit and approximate the real data distribution without supervision. The use of generators and discriminators for confrontation makes the fusion image retain richer information, and the end-to-end network structure no longer needs to manually design fusion rules.

2.1. Information Fusion Network Framework. The generative confrontation network was proposed by Goodfellow in 2014 [5] and is widely used in the field of deep learning. Generative adversarial network is a two-person zero-sum game idea, which can effectively estimate the distribution of data characteristics and generate new samples. The generative confrontation network includes a generative model (G) and an identification model (D). The generative model has the ability to fit the distribution of image data, and the discrimination model can estimate the probability that the input sample is real data. The purpose of the generator is to generate sample data. The sample data distribution is P_z. The training process of generating a confrontation network is to make the data distribution P_z of the generated data infinitely close to the real data distribution P_r. The specific formula is as follows.

$$\min_G \max_D V_{\mathrm{GAN}}(D,G) = \mathrm{E}_{x\sim p_r}[\log D(x)] + \mathrm{E}_{z\sim p_z}[\log(1 - D(G(z)))]. \tag{1}$$

It can be seen from the above formula that PG cannot show that if the discriminator is trained too well or too poorly, the generator will not get effective gradient descent, and the two cannot be updated synchronously, which will cause the GAN training to collapse. To solve this problem, the solution is to make the discriminator meet the Lipschitz [6] continuity condition (Lipschitz continuity):

$$\left|f(x_1) - f(x_2)\right| \le K\left|x_1 - x_2\right|. \tag{2}$$

For f, the smallest constant K is called the Lipsch constant of f. Limit the gradient of the discriminator to a certain range, so that the discriminator can gradually update the gradient in a small range.

This paper establishes a dual discriminator GAN for multimodal image fusion, and the overall framework is shown in Figure 1.

The generator in the figure above represents the generator of the fusion image, the input channel is connected with the visible light image and the infrared image, and the fusion image is input to the discriminator in a single-channel manner. The dual discriminators discriminator I and discriminator V are used to distinguish between fusion image and infrared image and fusion image and visible light image, respectively. After the continuous confrontation and iterative update between the generator and the discriminator, the trained generator is obtained. Single channel represents the single channel that contains the source image and the fusion image when the input of each discriminator is input. If the input contains both the fusion image and the corresponding source image as the dual channel of conditional information, the task of the discriminator will be simplified to whether the input image is same. This is too simple for the discriminatory network, and it is impossible to establish an effective confrontational relationship between the generator and the discriminator.

2.1.1. Generator Network Structure. The generator contains a total of six convolution modules, each of which contains a convolution layer and an activation function and uses the same $3 * 3$ convolution kernel. The number of convolution kernels for the first 5 convolution modules is 32. This can ensure that the network structure fully extracts image features. The generator structure diagram is shown in Figure 2.

2.1.2. Discriminator Network Structure. The discriminator contains a total of 6 convolution modules. Each convolution module contains a convolution layer and an activation function. The convolution size is set to 3, and the number of convolution kernels is set to 64, 128, and 256, respectively. Finally, it contains two fully connected layers. The discriminator structure diagram is shown in Figure 3.

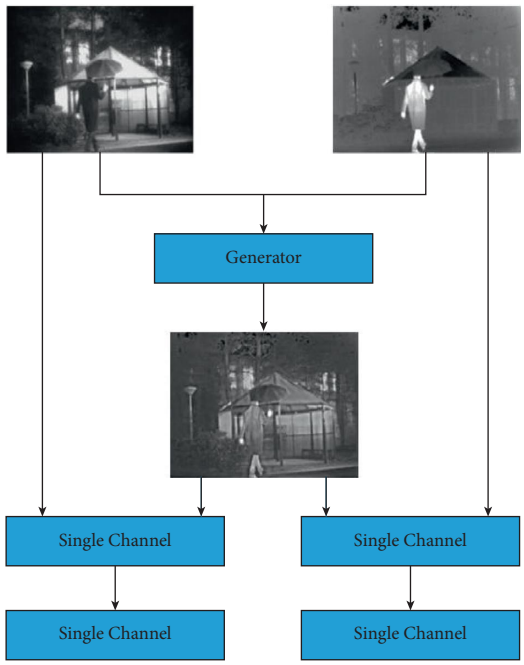


Figure 1: Image fusion based on GAN.

2.1.3. Loss Function

(1) Generator Loss Function. The generator loss function is defined as follows:

$$L = L_{\text{advers}}(G) + \lambda_1 L_{\text{content}}. \tag{3}$$

L is the total loss of the generator, $L_{\text{advers}}(G)$ represents the confrontation loss, L_{content} represents the content loss, and 1 is the coefficient. In order to make the generated image, save the infrared and visible light information as much as possible, the content loss of the generator is defined as L_{content}.

$$L_{\text{content}} = \frac{1}{HW}\left(\mu\left\|I_f - I_r\right\|_F^2 + \gamma\left\|\text{LBP}\left(I_f\right) - \text{LBP}\left(I_v\right)\right\|_F^2\right). \tag{4}$$

Here, μ and γ are the coefficients, H and W are the length and width of the input image, and I_f, I_r, and I_v are the fusion image, infrared image, and visible light image. The first item in the bracket is for the fusion image to save more information from the infrared image, and the second LBP function is defined as shown in formula (5), the purpose is to make the fusion image save more texture information from the visible light image.

$$\text{LBP}\left(x_c, y_c\right) = \sum_{p=0}^{p-1} 2^p s\left(i_p - i_c\right). \tag{5}$$

Here, (x_c, y_c) is the central pixel, and its pixel intensity value is i_c, i_p. $L_{\text{advers}}(G)$ stands for confrontation loss. The confrontation loss consists of two parts: the confrontation loss between the generator and the discriminator 1 and the confrontation loss between the generator and the discriminator 2, and the definition is shown as follows:

$$L_{\text{advers}}(G) = -\sum_{i=1}^{N} \text{E}_{z\sim p_g}\left[D_i(z)\right], \tag{6}$$

where z represents the generated data, p_g represents the distribution of the generated data, and N represents the number of discriminators (N takes 2).

(2) Discriminator Loss Function. Although the fusion image generated by the generator can save infrared and visible light information to a certain extent, it still needs to use the generated image and the source image to save more detailed information through the discriminator. The discriminator loss function is shown as follows:

$$L_{D_{ir}} = -\text{E}_{x\sim p_{ir}}\left[D_{ir}(x)\right] + \text{E}_{z\sim p_g}\left[D_{\text{vis}}(z)\right] + \lambda_3 \text{E}_{\tilde{x}}\left[\left(\left\|\nabla_{\tilde{x}} D_{ir}(\tilde{x})\right\|_2 - 1\right)\right]. \tag{7}$$

Here, $L_{D_{ir}}$ represents the loss of the visible light image and the generated fusion image as the input of the discriminator, p_{vis} and p_g represent the visible light image distribution and the distribution of the generated image, and λ_3 is the hyperparameter.

2.2. Improved Target Detection Algorithm. The target detection task based on the deep convolutional neural network has made great progress with the rapid development of deep learning, and the detection accuracy has been significantly improved compared with traditional detection methods. Many scholars have designed many detection networks. In general, the detection network is roughly divided into two-stage target detection and single-stage target detection. The two-stage target detection network has a candidate frame extraction step. Compared with the single-stage, the accuracy is higher, but the network prediction speed is slower. From R-CNN [7] to faster R-CNN [8], network detection accuracy is getting higher and higher, and the detection speed is getting faster and faster. Faster R-CNN is a classic structure in a two-stage target detection network. The network structure diagram is shown in Figure 4.

The faster R-CNN target detection algorithm is to define convolution feature extraction, candidate frame selection, candidate frame classification, and bounding box regression in a network, which can be regarded as faster R-CNN is the RPN (region proposal network) network and fast R—the combination of CNN network and the convolutional layer of RPN [9] is shared with fast R-CNN. The specific method is shown in Figure 5.

Many scholars have successively proposed improvement strategies for deep convolutional neural networks, some articles have improved the loss function, and some have proposed new improvement ideas such as deformable convolution and expanded convolution. This article focuses on the improvement ideas of the nonmaximum supression (NMS) [10] algorithm. The function of the NMS algorithm is to remove redundant detection results, and only keep a bounding box as the output of the detection result, which has very important significance for target detection. The original detection results of the network often produce multiple bounding boxes near the same target. At this time, it is necessary to sort according to the probability value

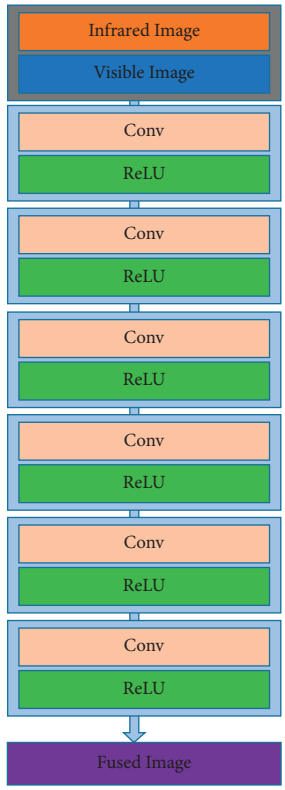


Figure 2: Generator network structure.

of the bounding box classification and select the bounding box with the highest score as the final detection result of the target at that location. If the remaining bounding boxes are such that if the IoU value of the selected bounding box is greater than the set threshold, it will be eliminated directly. The NMS algorithm is shown as follows:

$$S_i = \begin{cases} s_i, & \text{IoU} < N_t, \\ 0, & \text{IoU} \geq N_t. \end{cases} \tag{8}$$

The disadvantage of the NMS algorithm is that if the two targets on the image are relatively close, the IoU value of the bounding box between the target and the target is very large, which is very easy to cause a target to be undetected. Therefore, in view of the above shortcomings, this paper adopts the soft-NMS algorithm. The purpose is to select a bounding box with the highest current score and then update the score according to the IoU [11] value between the surrounding bounding box and the boundary with the highest score. A bounding box with a larger IoU value has a lower update score; a bounding box with a not too large IoU value will not have a too low score after the update, so that problems caused by NMS can be avoided to a certain extent. Soft-NMS is shown as follows:

$$S_i = \begin{cases} S_i, & \text{IoU} < N_t, \\ S_i e^{(\text{IoU}^2/\delta)}, & \text{IoU} < N_t. \end{cases} \tag{9}$$

2.3. Multimodal Information Fusion Detection. The algorithm in this paper regards the entire image fusion process as a process of confrontation between the generator and the discriminator. The training of the network model is not exactly the same as the test. Only the trained generator is needed during the test, and no discriminator is required to participate. In the confrontation generation network of the double discriminator, the first discriminator is mainly used to discriminate infrared images and generate images, and the other discriminator discriminates visible light images and generates images. The purpose is to enable the generated images to save infrared image temperature information and visible light gradient information, to avoid problems such as insufficient storage of single discriminator information and rely on the confrontation generation network to map visible light image information and infrared image information to the same feature space. At this time, the target detection task is similar to the visible light target detection. The feature extraction network and the classification network are completed. The detection framework is shown in Figure 6.

3. Experimental Results and Analysis

The experimental environment is configured as Ubuntu16.04 operating system, Pytorch deep learning framework; the hardware environment is NVIDIA GTX 1080ti graphics card ×2, Intel Core i7 processor. The experimental part uses FLIR [12] infrared data set for algorithm verification. The data set has two domains: infrared domain and visible light domain. The infrared image contains 7153 images, and the visible light image contains 6936 images. The detection categories are divided into three categories: people, cars, and bicycles.

3.1. Fusion Experiment. The fusion method in this paper is analyzed and compared with other fusion algorithm methods.

3.1.1. Qualitative Evaluation. The experimental results show that the fusion algorithm used in this paper has richer background detail information, such as the sky information in the two images, which obviously saves more texture information. In addition, compared with the single discriminator FusionGAN [13] algorithm, it can obviously get a better fusion effect, and it can reflect the prominent target and detailed features of the source image better, which is helpful for the next target detection. The fusion effect of different algorithms is shown in Figure 7.

3.1.2. Quantitative Evaluation. It mainly uses quantitative evaluation index fusion methods such as information entropy (EN), standard deviation (SD), mutual information (MI), and peak signal-to-noise ratio (PSNR) [14].

It can be seen from Figure 8 that the fusion algorithm in this paper has achieved obvious advantages in the three indicators of MI, EN, and SD, especially in the SD indicator, which shows great superiority. This can reflect to a certain extent that the fusion algorithm in this paper not

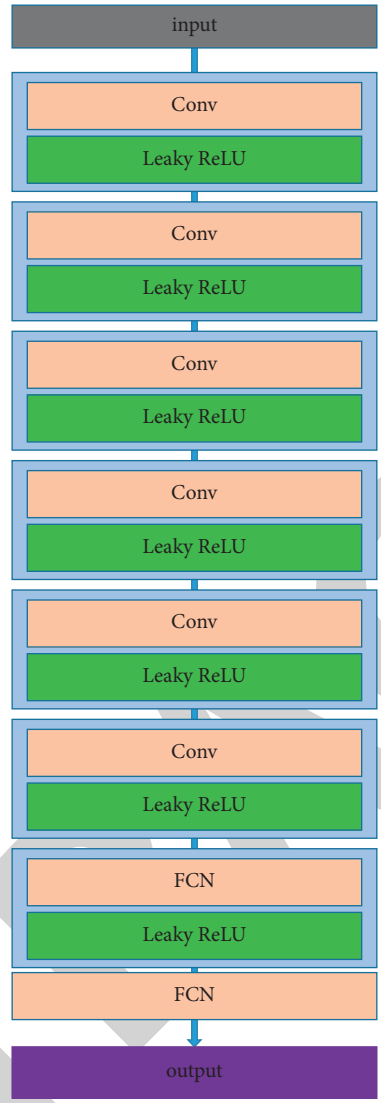


Figure 3: Discriminator network structure.

only has better visual effects but also has obvious advantages in quantitative evaluation.

3.2. Target Detection Model

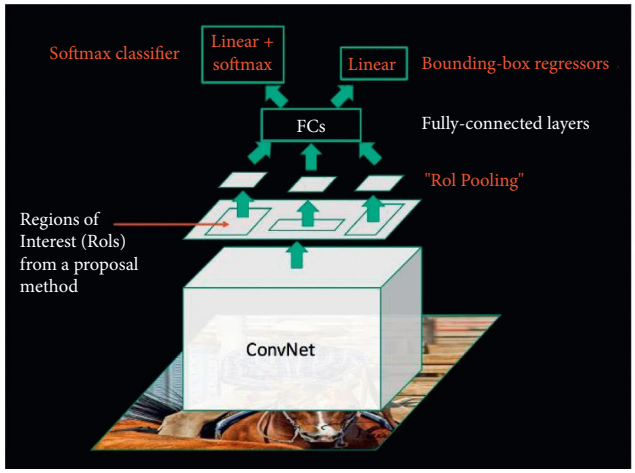
3.2.1. Target Detection Model Training. First, use the abovementioned trained generator to fuse visible light and infrared images to obtain a fusion image containing multimodal information and then use the image data set to train the fusion image target detection model.

After 50,000 iterations of training, the training loss of the visible light target detection model is about 0.5, and its loss transformation curve is shown in Figure 9.

The change curve of the intersection ratio between the predicted bounding box and the actual bounding box is shown in Figure 10. The abscissa represents the number of iterations, and the ordinate represents the intersection ratio of the predicted bounding box and the actual bounding box. As the number of iterations increases, the intersection ratio becomes the overall. The upward trend is finally close to 1, which means that the predicted bounding box in the visible light scene is very close to the actual bounding box, which meets the training requirements.

3.2.2. Target Detection Experiment. The target detection model generally uses mAP (mean average precision) [15, 16] index for evaluation, which is the average value of the average detection accuracy (average precision, AP) of multiple types of objects. The test sets under the visible light and infrared scenes were tested, respectively, and the results are shown in Tables 1–3.

It can be seen from the above table that the method in this paper has a high accuracy rate in the overall structure, can effectively fuse the bimodal information, and realize the accurate description of the scene information. The model detection effect is shown in Figure 11.



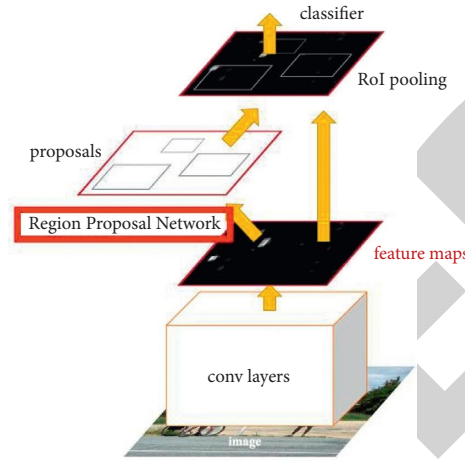


Figure 4: Faster RCNN network structure.

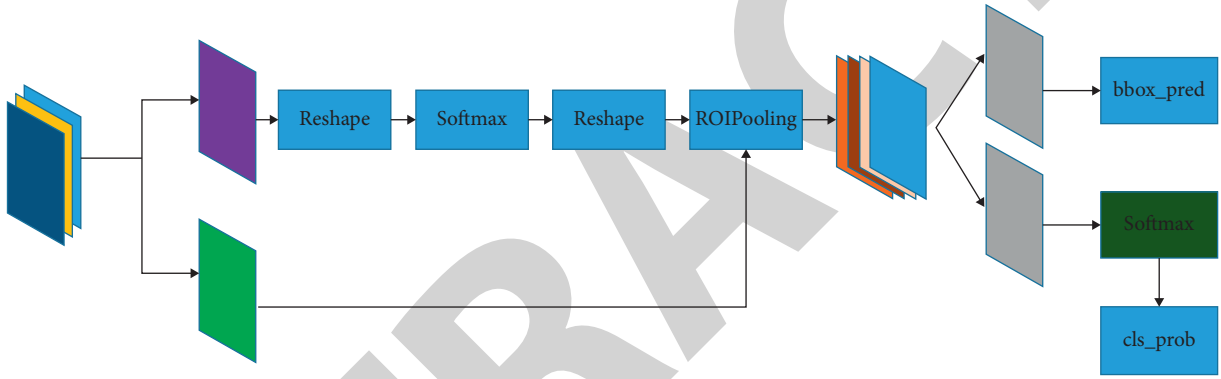


Figure 5: RPN network structure.

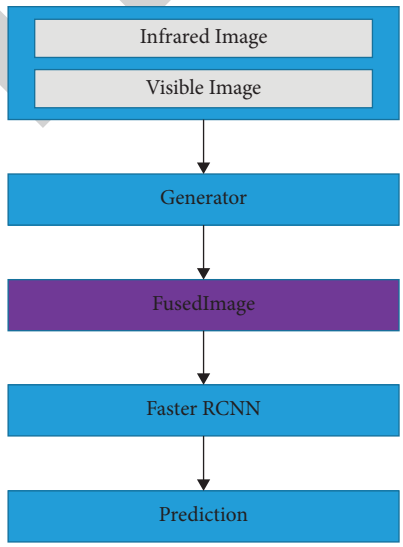


Figure 6: Multimodal information fusion detection algorithm.

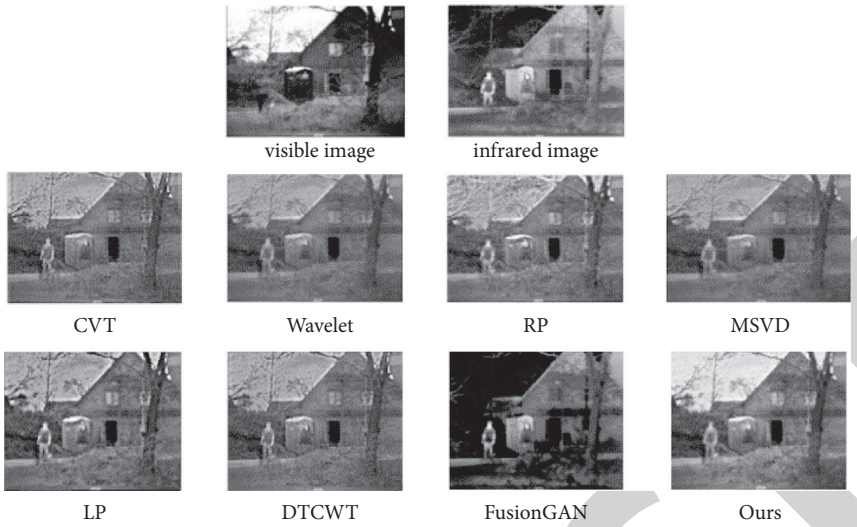


Figure 7: Comparison of the effect of different fusion algorithms.

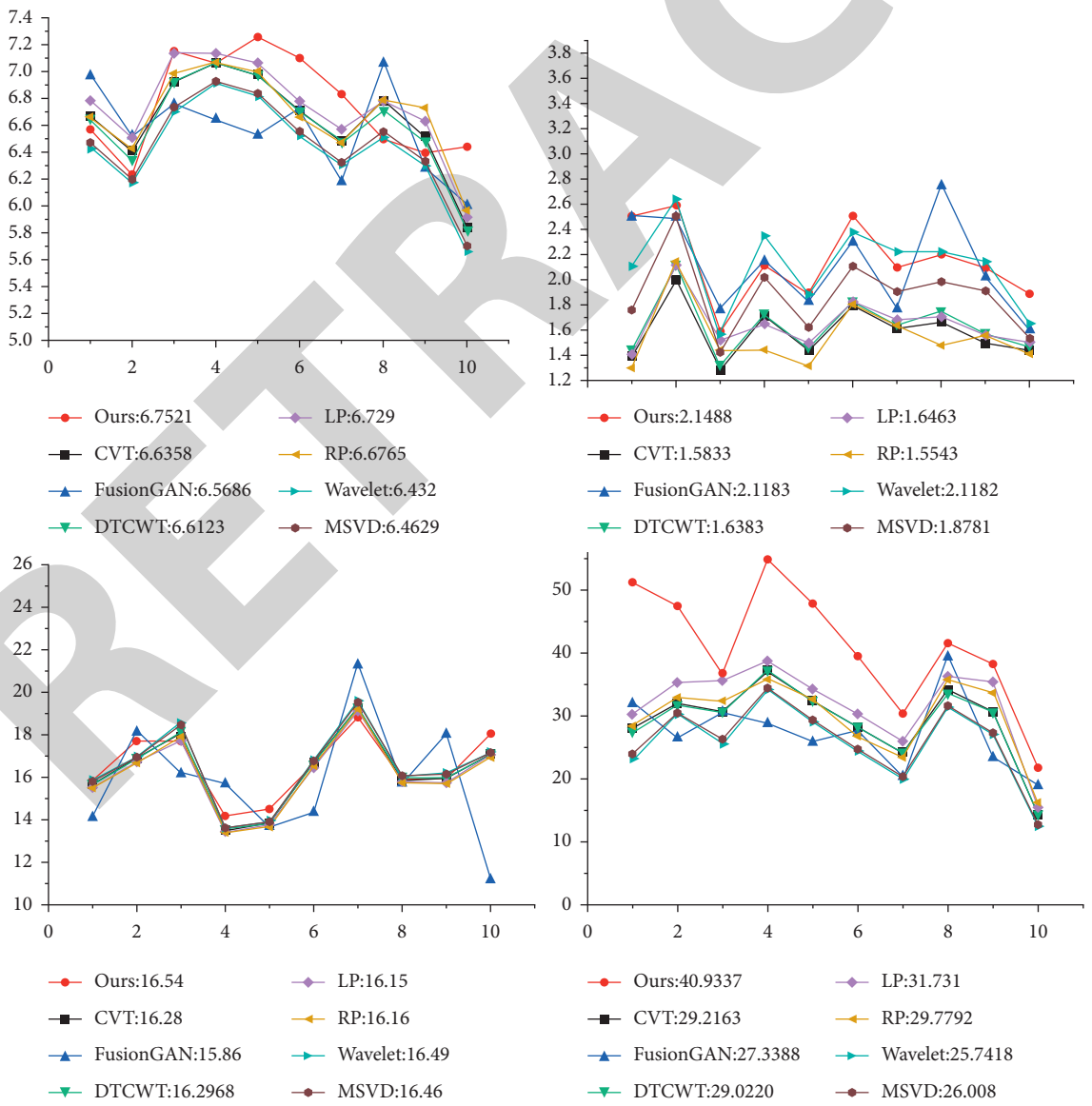


Figure 8: Quantitative evaluation.

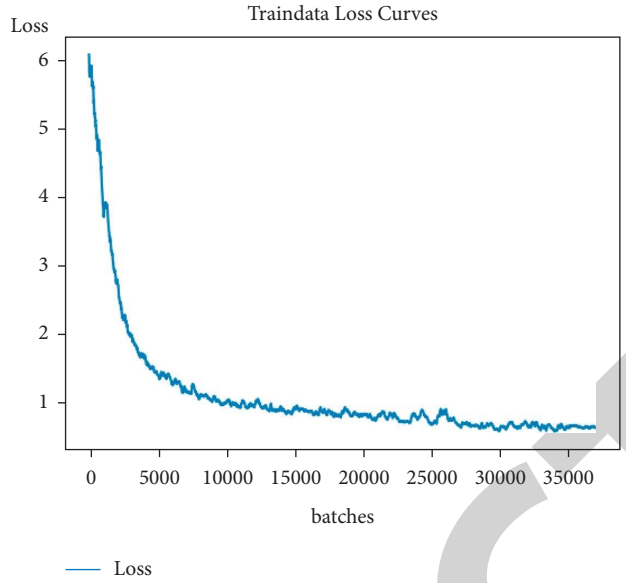


Figure 9: Training loss curve of detection model.

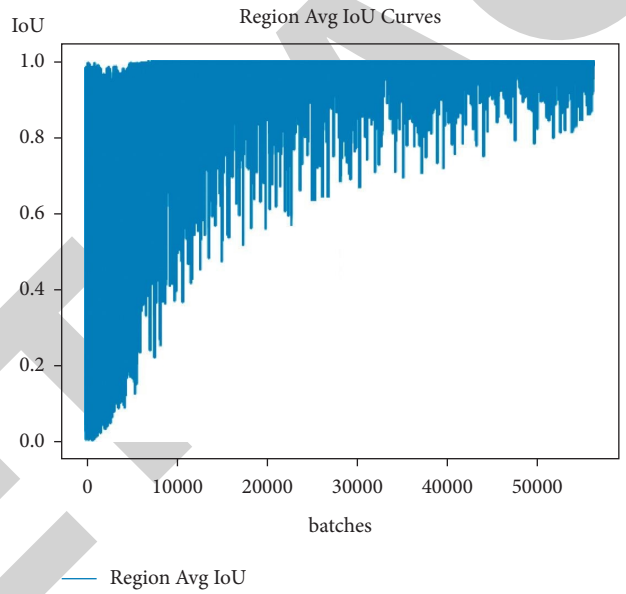


Figure 10: Curve of change of intersection ratio of detection model.

Table 1: Visible.

Model	AP (%)
Faster RCNN (+NMS)	86.28
Faster RCNN (+soft-NMS)	89.35
Ours	93.82

Table 2: Infrared.

Model	AP (%)
Faster RCNN (+NMS)	84.39
Faster RCNN (+soft-NMS)	87.16
Ours	93.36

Table 3: Fused.

Model	AP (%)
Faster RCNN (+NMS)	74.39
Faster RCNN (+soft-NMS)	79.16
Ours	95.36

Figure 11: Fusion model detection effect diagram.

4. Conclusion

Multimodal information fusion has a wide range of application scenarios. This paper designs a confrontation generation network that can realize end-to-end training to fuse multimodal information to improve the complementarity and low redundancy among multimodal information features and improve the accuracy of target detection and classification based on fusion features. Multimodal information fusion provides richer target information than single-modality, but it also greatly increases the amount of calculation, which makes it difficult to achieve real-time detection effects in application scenarios with limited computing resources.

Data Availability

The simulation experiment data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Development and Application of Identification Control System in Epidemic Prevention and Control Area (Grant nos. JYJFZX20-01) and National Educational Technology Research Project of Central Audio Visual Education Center (Grant no. 186130061). Xuzhou city will promote the special Key Research and Development Plan for Scientific and Technological Innovation (industrial key technology research and development) project "R&D and Application of Water Resources Cloud Control Platform at River Basin Level (Grant no. KC21108), special policy guidance plan for scientific and technological innovation (industry-university-research cooperation) Big Data-Based Multi-Objective Coordinated and Balanced Allocation of Large-Region Water Resources (Grant no. KC21335), school level mixed teaching team of computer network technology specialty group by the Academic Affairs Office of Jiangsu Construction Vocational and Technical College (Grant no. jw2021-8), and the research and practice of demonstration vocational education group—Taking Xuzhou Huaihai Service Outsourcing Vocational Education Group as an example (Grant no. ES2021-2).

References

[1] Z. Wang, "A new approach for segmentation and quantification of cells or nanoparticles," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 3, pp. 962–971, 2016.

[2] Z. Wang, J. Xiong, Y. Yang, and H. Li, "A flexible and robust threshold selection method," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2220–2232, Sept. 2018.

[3] T. Mantecón, C. R. del Blanco, F. Jaureguizar, and N. García, "Hand gesture recognition using infrared imagery provided by leap motion controller," in *Proceedings of the International Conference on Advanced Concepts for Intelligent Vision Systems, ACIVS 2016*, pp. 47–57, Lecce, Italy, October 2016.

[4] E. Persoon and K.-S. Fu, "Shape discrimination using Fourier descriptors," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 7, no. 3, pp. 170–179, 1977.

[5] C. Li, H. Cheng, S. Hu, X. Liu, J. Tang, and L. Lin, "Learning collaborative sparse representation for grayscale-thermal

tracking," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5743–5756, 2016.

[6] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, "RGB-T salient object detection via fusing multi-level CNN features," *IEEE Transactions on Image Processing*, vol. 29, pp. 3321–3335, 2020.

[7] Z. Zhang, Z. Lin, J. Xu, W.-D. Jin, S.-P. Lu, and D.-P. Fan, "Bilateral attention network for rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 1949–1961, 2021.

[8] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, "Rgbd salient object detection via deep fusion," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2274–2285, 2017.

[9] H. Song, Z. Liu, H. Du, G. Sun, O. Le Meur, and T. Ren, "Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4204–4216, 2017.

[10] D. Guan, Y. Cao, J. Yang, Y. Cao, and M. Y. Yang, "Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection," *Information Fusion*, vol. 50, pp. 148–157, 2019.

[11] H. Nam and B. Han, "Learning multidomain convolutional neural networks for visual tracking," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4293–4302, Las Vegas, NV, USA, June 2016.

[12] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proceedings of the 2017 IEEE conference on computer vision and pattern recognition*, pp. 472–480, Honolulu, HI, USA, July 2017.

[13] L. Liangkui, W. Shaoyou, and T. Zhongxing, "Using deep learning to detect small targets in infrared oversampling images," *Journal of Systems Engineering and Electronics*, vol. 29, no. 5, pp. 947–952, 2018.

[14] C. Li, D. Song, R. Tong, and M. Tang, "Illumination-aware faster R-CNN for robust multispectral pedestrian detection," *Pattern Recognition*, vol. 85, pp. 161–171, 2019.

[15] M. Silvagni, A. Tonoli, E. Zenerino, and M. Chiaberge, "Multipurpose UAV for search and rescue operations in mountain avalanche events," *Geomatics, Natural Hazards and Risk*, vol. 8, no. 1, pp. 18–33, 2017.

[16] Z. Wu, X. Wang, and C. Chen, "Research on lightweight infrared pedestrian detection model algorithm for embedded Platform," *Security and Communication Networks*, vol. 2021, Article ID 1549772, 7 pages, 2021.