Visible and Infrared Image Fusion for Object Detection: A Survey

Yuxuan Sun[1], Yuanqin Meng[1], Qingbo Wang[2], Minghua Tang[1], Tao Shen[1], and Qingwang Wang[1](✉)

[1] School of Information Engineering and Automation, Yunnan Key Laboratory of Computer Technologies Application, Kunming University of Science and Technology, Kunming, China
wangqingwang@kust.edu.cn
[2] Naval Research Academy, Beijing, China

Abstract. Thanks to the advancements of Deep Learning (DL) Algorithms. DL-based object detection models has witnessed remarkable success in the past few years. By leveraging deep convolutional neural networks (CNNs) and other deep learning models, rich feature representations can be effectively extracted from visible (RGB) images for object detection. However, for challenging scenarios such as low-light conditions and haze, the performance of object detection using visible images is often not satisfactory. On the other hand, thermal camera, which is unaffected by lighting conditions, can penetrate through low-light and hazy environments to capture object images. However, infrared images lack edge and texture information of objects. Recognizing the complementary nature of visible and infrared images, researchers have explored the fusion of visible and infrared images for object detection, yielding promising research outcomes. This paper provides an analysis of the current research status of visible and infrared image fusion for object detection. Firstly, the fusion models are categorized into three kinds: pixel-level fusion, feature-level fusion, and decision-level fusion. Several models within each category are discussed. Furthermore, this paper summarizes five datasets that can be utilized for training RGB-Infrared object detection models and compares the experimental results of selected models on the KAIST dataset. Lastly, the paper concludes with a summary of existing challenges in the field and offers some reflections on future directions.

Keywords: object detection · multimodal fusion · RGB-Infrared image · deep learning · cross modal

1 Introduction

In the field of computer vision, object detection is considered as one of the most fundamental tasks. And it has extensive practical applications in domains such as firefighting, military operations, and autonomous driving. Therefore, the research on object detection techniques holds significant practical significance. Notable advancements have been made in recent years in object detection techniques, particularly those that utilize deep learning methods [1]. Nonetheless, the performance of single-modal object detection

P. You et al. (Eds.): ICIVIS 2023, LNEE 1163, pp. 236–248, 2024.
https://doi.org/10.1007/978-981-97-0855-0_24

models that trained on visible (RGB) image dataset is not satisfactory in scenarios with low light conditions, haze, or abnormal lighting due to the substantial influence of environmental factors on visible light. Consequently, scholars have started to consider the synergistic utilization of light sources from other spectral bands in conjunction with visible light for object detection.

Infrared light, as a non-visible form of light, possesses the ability to penetrate fog and smoke, remaining unaffected by lighting conditions, thus enabling object detection in adverse environments. However, photographs captured using infrared light have low resolution and contain limited texture information, making it challenging to provide detailed object information. On the other hand, visible light offers rich information and high resolution but exhibits suboptimal detection performance under environmental influences. With the goal of enhancing the accuracy and dependability of object detection, scholars have explored the methods for fusion of infrared and visible images, [2, 3]. Experiments have yielded promising results, confirming the feasibility of this approach in harnessing the distinct strengths of both modalities to overcome their individual limitations. Currently, several datasets, such as KAIST [4], CVC-14 [5] and M^3FD [2], are available for training RGB-Infrared object detection models.

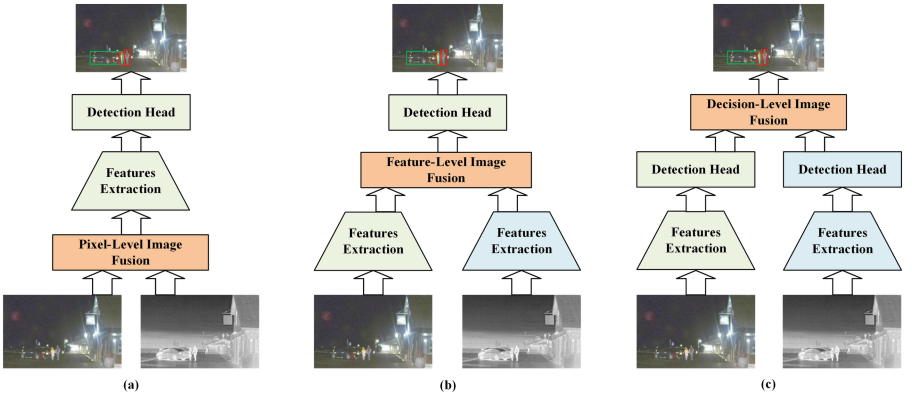

Fig. 1. Three ways of fusing visible and infrared images in multimodal object detection task: (a) show pixel-level fusion, (b) show feature-level fusion, (c) show decision-level fusion.

In recent years, thanks to the rapid development of multi-modal models and the application requirements of object detection in real life scenarios, and considering the significant information complementarity between infrared and visible light, research on object detection based on the fusion of visible and infrared images remains a topic of considerable interest. Therefore, there is a need for a comprehensive review paper to summarize the recent advancements in RGB-Infrared models and discuss the existing challenges, thereby providing assistance to researchers in the field. The paper's three main contributions can be summarized as follows:

- The paper categorizes fusion methods for RGB-Infrared object detection into three types: pixel-level fusion, feature-level fusion, and decision-level fusion. It also presents several representative models for each category..

- A comprehensive comparison of RGB-Infrared object detection models is conducted on KAIST dataset, providing abundant comparative analysis.
- This review provides a summary of the current challenges in RGB-Infrared object detection and offers valuable insights into future research directions for scholars in the field.

The following sections of this review will be structured as follows: Sect. 2 provides a review of fusion methods and introduces several representative models. Section 3 discusses existing datasets for training RGB-Infrared object detection models. Section 4 presents a performance comparison of several models. Section 5 summarizes the current challenges and offers thoughts on future research directions in the field. Finally, Sect. 6 gives a conclusion.

2 Rgb-Infrared Object Detection

This section will provide an explanation of three fusion strategies used for multimodal object detection, which are: pixel-level fusion, feature-level fusion, and decision-level fusion, the schematics of the three fusion strategies are shown in Fig. 1. We will present several representative models for each fusion approach. In this section, symbol V and T denote the original visible and infrared image.

2.1 Pixel-Level Fusion

Pixel-level fusion involves integrating the information of each pixel in infrared and visible images. This method operates directly at the pixel level, combining the pixel-level details from both types of images to extract a broader range of target information.

Pixel-level fusion typically involves various computational procedures and manipulations, including channel concatenation [6], weighted averaging [7], and channel substitution [8]. The most mainly advantage of pixel-level fusion lies in its capacity to fully exploit all the information embedded within the images, beyond the extraction of specific details from other levels (such as region or feature levels).However, although pixel-level fusion effectively utilizes all the information present in the images, it may encounter challenges such as image misalignment [7], the incorporation of redundant information, and the introduction of noise. Therefore, to optimize fusion outcomes and enhance image performance and accuracy, techniques like noise suppression, image sharpness enhancement, and image alignment are required.

2.2 Wagner's Work

The model proposed in [6] is the first of its kind to fuse visible and infrared images for deep learning-based object detection tasks. The authors argue that object detection models trained solely on visible images perform poorly at night and are susceptible to lighting conditions. Therefore, there is potential for enhancing object detection performance by integrating information from visible and infrared images. Building on this idea, the authors put forward two multimodal image fusion techniques: early fusion and

late fusion. The early fusion strategy, which is employed to integrate the information from visible and infrared images, falls under the purview of pixel-level fusion discussed in this paper.

The network structure using early fusion strategy is shown in Fig. 2, the authors adopt a direct channel-wise concatenation approach to merge the RGB and infrared images at the input stage, forming a 4-channel input image. This fused image is then fed into a modified CaffeNet to extract features, which are subsequently utilized as inputs for the subsequent network for object detection task. By performing pixel-level fusion of the two modalities at the input stage, the authors aim to enhance the utilization of inter-modality correlations for improved object detection.

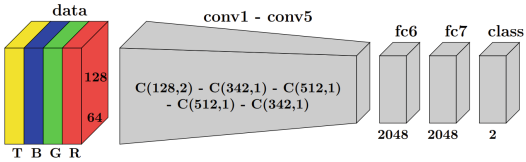

Fig. 2. Early fusion (pixel-level fusion) at input end and the Figure is from [6].

Despite efforts, the experimental results have not been satisfactory. The performance of the multi-modal object detection model, employing the early fusion strategy, on the KAIST data set is inferior to that of the model trained solely on pure RGB images. Even trained the early fusion model with pre-training, The final performance of the model is still slightly inferior to that of the earlier multimodal target detection method ACF + T + THOG [4] used by the authors to compare.

1) *Vandersteegen's Work*

Vandersteegen proposed thress method for pixel-level fusion of multimodal images at the input stage based on weighted averaging and channel substitution in the literature [8].

Based on the viewpoint that infrared images may contain more information than any individual color channel in RGB images in certain cases, the author presented the first implementation approach for pixel-level fusion. This approach replicated the RGB image three times and replacing one color channel in each replicated image with the corresponding channel from the infrared image, while keeping the other two color channels unchanged. The resulting three images were named TGB, RTB, and RGT, with T representing the specific color channel that was replaced in the RGB image.

However, directly replacing the color channels in the original RGB image may result in significant information loss. Therefore, the author proposed a second fusion strategy which transformed the RGB image into a representation with imbalanced inter-channel information. Subsequently, the infrared image is used to replace the less important channel in the transformed representation, thereby preventing excessive information loss during pixel-level fusion of multimodal images. In this regard, the author adopts a specific approach. Initially, the RGB image is transformed into the LUV representation.

Subsequently, one of the U or V channels is replaced with the corresponding channel from the infrared image, yielding the fused image.

Although the second fusion strategy which transforms the image representation before replacement, can reduce information loss, the elimination of one channel's image information will still have an impact on the model's performance. Hence, in the third fusion strategy, the author opts not to exclude any channel from the input images. Instead, a weighted average technique is employed to fuse the three channels of the RGB image with the infrared image. The process can be represented as follows [8]:

$$X = \frac{R+T}{2} \; Y = \frac{G+T}{2} \; Z = \frac{B+T}{2} \tag{1}$$

where R, G, and B represent the three channels of the RGB image, T represents the infrared image, and X, Y, and Z represent the three channels of the fused image. The math operations are all performed at the pixel-level.

2.3 Feature-Level Fusion

Deep learning effectively leverages the strong feature extraction capabilities of deep neural networks, enabling a fusion of feature information from both infrared and visible images at the feature level. Convolutional neural networks (CNNs) are commonly employed to extract features from these images for the fusion process. This approach ensures that the combined features capture relevant information from both modalities, leading to enhanced performance in various tasks. This process involves feeding the infrared and visible images into either a single CNN or two separate CNN networks. By applying forward propagation to the network, two distinct sets of feature maps are generated. These feature maps are capable of capturing texture, shape, and edges of the images. In some models, the extracted feature maps are further processed through an Attention-Based module before fusion, in order to enhance the feature representation or suppress noise information [3, 9].

Common fusion methods for the fusion of the two obtained feature maps include feature concatenation [10], and feature weighting [9, 11]. Feature concatenation merges the two feature maps by channel concatenation to conduct a new fusion feature map. By performing weighted summation, feature weighting combines the two feature maps to create a new feature map. This fused feature map incorporates information from both infrared and visible light sources. Subsequently, the fused feature map is fed into subsequent networks for further processing, ultimately leading to the generation of object detection results.

2) *BAANet*

BAANet [9] is an exemplary model that utilizes feature-level fusion to merge visible and infrared images in order to facilitate object detection. At the core of BAANet lies the Bi-directional Adaptive Attention Gate (BAA-Gate), which effectively mitigates noise within each modality and maximizes the exploitation of inter-modal correlations, thereby enhancing multimodal image fusion.

The BAA-Gate operates in two distinct stages: the channel distilling process and the spatial aggregation process. In the first stage, it is tasked with performing image

denoising and recalibration. To achieve this, the visible and infrared images are initially combined along the channel dimension. Subsequently, a global average pooling layer is employed to extract the channel attention information from the images. Subsequently, a multilayer perceptron is used to compute the vector of channel distilling weights, which has the same length as the number of channels in the input images. These weights are then applied to the input images through channel-wise multiplication, resulting in distilled features. The distilled features are added to the input feature maps to achieve feature map recalibration. Additionally, considering the impact of changes in lighting conditions on the model's recognition performance, the authors have designed a recalibration operation with illumination-based weights based on the aforementioned recalibration operation.

The spatial aggregation process follows the channel distilling process. In this stage, the recalibrated multimodal images are concatenated along the channel dimension. Then, a 1×1 convolution operation is applied to obtain the spatial aggregation gate W_{Ts}. . The final fused image is obtained using (2) [9], where w_T, w_V denote the illumination weights generated by illumination-based sub-network, T_{rec} and V_{rec} denote the recalibrated visible and infrared images.

$$\begin{aligned} V_{fuse} &= V + w_T * (w_{Ts} \odot T_{rec}) \\ T_{fuse} &= T + w_V * (w_{Ts} \odot V_{rec}) \end{aligned} \tag{2}$$

This work is based on the bi-directional adaptive attention gate and achieves denoising, recalibration, fusion of multimodal images, and demonstrates excellent performance on the object detection task.

3) *TINet*

Leveraging the complementary information between different modalities for feature-level fusion is a crucial factor influencing the performance of multimodal object detection models. Based on this viewpoint, Zhang [11] proposed TINet, which adaptively fuses features from multimodal images and applies them to object detection.

TINet adopts several modules to achieve the fusion of multimodal images: the illumination-guided feature weighting (IGFW) module, the inter-modality attention (Inter-MA) module, and the intra-modality attention (Intra-MA) module. These modules work together to combine the information from different modalities effectively. The IGFW module generates illumination weight based on the original multimodal image's lighting conditions, which using for subsequent image fusion.

The original images are first passed through a backbone network, and the resulting feature maps are inputted to the Inter-MA module and Intra-MA module for processing. In the Inter-MA module, the feature maps are initially fed through a CNN network, and a weight vector D is obtained through global average pooling. Then, the first fusion of multimodal features is performed, as represented by the following [11]:

$$\begin{aligned} V_{inter} &= V + \mathcal{F}(V + (D \odot T)) \\ T_{inter} &= T + \mathcal{F}(T + (D \odot V)) \end{aligned} \tag{3}$$

where $\mathcal{F}$ represents the residual block. $\odot$ denotes the channel-wise multiplication.

In the Intra-MA module, the first step involves enhancing the intra-modal feature information of the images. This process can be represented as follows [11]:

$$\begin{aligned} V_{intra} &= V \otimes (1 + M_V) \\ T_{intra} &= T \otimes (1 + M_T) \end{aligned} \tag{4}$$

where M_R and M_T denote the heatmaps which generated by the combination of 1×1 convolution layer and rectified linear unit (ReLU) function with the input of visible image and infrared image, and $\otimes$ denotes the pixel-wise multiplication. Afterward, the feature maps generated by the Intra-MA and the Inter-MA are fused to produce the ultimate fused image. This procedure can be represented as follows [11]:

$$F = w_V \cdot (V_{inter} + V_{intra}) + w_T \cdot (T_{inter} + T_{intra}) \tag{5}$$

where w_V and w_T denote the fusion weight generated by IGFW module.

2.4 Decision-Level Fusion

Decision-level fusion is a technique employed to merge infrared and visible images at the decision-making stage. This fusion process occurs after feature extraction and classification stages, and it usually involves utilizing multiple independent networks to process the infrared and visible images separately. The ultimate decision is reached by combining the outputs of these networks, which typically encompass confidence scores and object coordinates in object detection tasks.

Decision-level fusion methods typically involve weighted fusion [12, 13] and decision combination [14]. Weighted fusion involves summing the outputs of each network using weighted coefficients and making a decision based on the result. In the process of decision combination, the bounding boxes extracted from both the visible and infrared images are merged, and the bounding box with the highest confidence score for each object is retained as the final detection result. This method ensures that the most confident and accurate detection results from both modalities are taken into account for the final decision. Decision-level fusion enables the comprehensive utilization of decision information from infrared and visible light. By combining information from both modalities, it can effectively address the limitations of a single modality in certain scenarios and significantly boost the system's robustness and reliability. This integration of multimodal data empowers the system to achieve more accurate and comprehensive results, making it better suited for real-world applications.

4) *IATDNN*

Under different lighting conditions, such as daytime and nighttime, visible and infrared images exhibit distinct features. However, using weight coefficients to fuse decision information would introduce redundant information and affect the performance of the model. Therefore, literature [15] designed a detection network called IATDNN, which learns fusion weights from the images and applies them to decision information fusion. In contrast to conventional decision-level fusion methods that directly fuse decision information extracted from visible and infrared images, IATDNN first performs

channel-wise fusion of infrared and visible light features during the feature extraction stage. These fused features are then fed into a sub-network called Conv-Pro to extract decision information under different lighting conditions. Finally, the decision information is fused based on learned weights.

There are two components for decision information fusion in IATDNN: illumination fully connected neural networks (IFCNN), which is used to generate weights for fusing decision information, and four sub-networks for extracting decision information. The initial visible and infrared images undergo processing through a convolutional neural network (CNN) to derive their respective feature maps. These feature maps, originating from both modalities, are then concatenated along the channel dimension, resulting in a fused feature representation, denoted as X. This fused feature X is subsequently fed into the two aforementioned components for further processing. The process of image processing in IFCNN can be expressed as

$$[w_d, w_n] = Softmax(f_2(f_{64}(f_{256}(P(X))))) \tag{6}$$

where w_d and w_n denote the decision fusion weights for daytime and nighttime illumination conditions, f_n denotes the fully connected layer which has n neurons, and P denotes the pooling layer.

The four subnetworks in IATDNN, named D-Cls, N-Cls, D-Bbox, and N-Bbox, which generate classification confidences c_d and c_n under daytime and nighttime illumination conditions, as well as bounding boxes b_d and b_n for both lighting conditions. Subsequently, these outputs are fused using the weights obtained from IFCNN to derive the final decision information. This fusion process can be described as follows [15]:

$$\begin{aligned} c_{fuse} &= w_d * c_d + w_n * c_n \\ b_{fuse} &= w_d * b_d + w_n * b_n \end{aligned} \tag{7}$$

5) IAF R-CNN

The literature [16] also takes into account the effect of illumination conditions on multi-modal object detection. However, in contrast to the literature [15], the authors of [16] argue that infrared images are not sensitive to lighting conditions. Therefore, they only utilize RGB images for estimating the illumination conditions and propose a multi-modal object detection model called IAF R-CNN, which implements the fusion of decision information based on illumination weightings.

In IAF R-CNN, the module responsible for estimating the illumination conditions of RGB images is called the Illumination-aware Network (IAN). IAN comprises two 3x3 convolutional layers, one ReLU layer, and a 2×2 max pooling layer. Afterward, there are two fully connected layers, one with 256 neurons and the other with 2 neurons. To address overfitting concerns, a dropout layer is included between these two fully connected layers, ensuring more robust and reliable estimation of the illumination conditions. The softmax values obtained from the network output serve as the estimated illumination values for the image.

The authors contend that under favorable lighting conditions, the RGB decision information should carry a higher weight, while the weight of infrared decision information should not be overly diminished. This balance ensures that the network can effectively leverage information from both modalities and benefit from their combined strengths.

However, in poor lighting conditions, the weight of infrared decision information should be higher, while the weight of RGB decision information should be lower, to prevent interference from irrelevant information in RGB modality. Based on this theory, the authors developed a Gated fusion strategy for decision information fusion. The process of calculating the fusion weights is described as follows [16]:

$$w = \frac{iv}{1 + a\exp(-\frac{iv-0.5}{b})} \tag{8}$$

where iv denotes the estimated illumination value generated by IAN which ranges [0, 1], a and b are parameters that can be learned

The decision information includes the classification confidence value s_V and bounding boxes' parameter b_V obtained from the RGB images, as well as the classification confidence value s_t and bounding boxes' parameter b_t obtained from the infrared images. The fusion process can be represented as follows [16]:

$$\begin{aligned} s_{fuse} &= w * s_V + (1 - w) * s_t \\ b_{fuse} &= w * b_V + (1 - w) * b_t \end{aligned} \tag{9}$$

Additionally, Table 1 summarizes the representative RGB-Infrared object detection models in recent years and classifies them according to fusion methods.

Table 1. Brief Summary of Fusion Categories and Key Related Works

Presentation Time	Fusion Strategy		
	Pixel-Level Fusion	*Feature-Level Fusion*	*Decision-Level Fusion*
Before 2021	Wagner's [6] French's [7] Vandersteegen's [8]	CFR [17]	IFCNN + IATDNN [15] IAF R-CNN [16]
2021	N/A	UFF + UCG [10]	IT-MN [12]
2022	N/A	TarDAL [2] RISNet [3] BAANet [9]	CMPD [13]
2023	N/A	TINet [11]	Hu's [14]

3 Datasets

Currently, there are several datasets available for training RGB-Infrared object detection models. In comparison to single-modal image datasets, multi-modal datasets consist of pairs of infrared-visible image combinations. The visible images are captured using conventional camera devices, providing visual information with red, green, and blue channels for each pixel. These images offer intuitive visual cues. On the other hand, infrared images are obtained using infrared cameras or infrared imaging devices, capturing the infrared radiation emitted by objects in the scene. Unlike visible-light images, infrared images depict the infrared distribution and heat characteristics of objects.

Table 2. Datesets for RGB-Infrared Object Detection Model Training

Dataset	*Year*	*Image Quantity*	*Image Resolution*	*annotated*	*aligned*
KAIST [4]	2015	95328	640 × 512	True	True
CVC-14 [5]	2016	8518	640 × 512	True	False
FLIR [18]	2018	10000	640 × 512	Only Infrared	False
LLVIP [19]	2021	30976	1080 × 720	True	True
M^3FD [2]	2022	8400	1024 × 768	True	True

The multi-modal datasets consist of image pairs acquired simultaneously in the same scene, ensuring consistent viewpoints between the visible and infrared images. Table 2 presents several datasets currently available for training RGB-Infrared object detection models.

4 Experiment

In this chapter, we select two RGB-Infrared target detection models for comparison from three fusion methods: pixel-level fusion, feature-level fusion and decision-level fusion, respectively. The evaluation is performed on the widely adopted KAIST dataset. The evaluation metric utilized is the log-average miss rate (MR^{-2}), which calculated through the false positive per image (FPPI) – miss rate (MR) graph.

MR means miss rate, which quantifies the rate of missed detections within the detection results. The computation process for MR is defined as follows:

$$MR = \frac{FP}{GT} = 1 - Recall \tag{10}$$

where *FP* denotes false positive, which indicates objects that are present in the image and annotated but not detected. *GT* represents ground truth, and the *MR* value can also be obtained by subtracting the model's Recall value from 1. *FPPI* means false positive per image, which is used to describe the average error rate per image, The calculation process can be expressed as follows

$$FPPI = \frac{FP}{N} \tag{11}$$

where N denotes the total number of the images.

To calculate the value of MR^{-2}, we first uniformly sample 9 points within the range of *FPPI* from 0.01 to 1. Then, we find the corresponding *MR* values for these 9 points as $m_1, m_2, \cdots m_9$, and the process that calculates MR^{-2} can be expressed as

$$MR^{-2} = exp\left(\frac{1}{9}\sum_{n=1}^{9} ln(m_n)\right) \tag{12}$$

The comparative results are presented in Table 3, where the experimental data are derived from the raw data provided in the respective papers.

Table 3. Comparison of Experimental Data Based on KAIST Dataset

Category	Method	Publication	MR^{-2}		
			All	*Day*	*Night*
Pixel-Level Fusion	Wagner's [6]	ESANN2016	53.94	50.90	51.76
	Vandersteegen's [8]	ICIAR2018	42.6	45.5	36.9
Feature-Level Fusion	BAANet [9]	ICRA2022	**7.92**	**8.37**	**6.98**
	TINet [11]	IEEE TIM 2023	9.15	10.25	7.48
Decision-Level Fusion	IATDNN [15]	Information Fusion 2019	26.37	27.29	24.41
	IT-MN [12]	IEEE TNSE 2022	14.01	14.10	13.87

5 Future Outlook

In recent years, scholars have constantly proposed models based on the fusion of visible light and infrared light for target detection in different ways, and achieved relatively excellent performance, but the performance of RGB-Infrared target detection model is still far from that of human eyes, so multi-modal target detection is still a direction that needs further research. This section will discuss the remaining challenges in the field of RGB-Infrared object detection, and give some thoughts on the future development direction.

5.1 Dataset Images Misalignment

Currently, the widely used dataset KAIST for training multi-modal object detection models consists of aligned visible-infrared image pairs in spatial domain. However, in practical applications, there exists pixel-level inconsistency between infrared and visible imaging processes. For example, the resolutions of infrared and visible imaging are usually different, leading to the inability to directly align the details and features in the images. Moreover, geometric distortions may occur due to device limitations or imaging environment influences. In addition, there are also alignment issues caused by scale disparities and visual angle differences. Currently, there is limited research on the alignment of visible and infrared images, and robust methods applicable to large-scale data are lacking. Therefore, further exploration is needed in this area.

5.2 Modal Image Instability

The issue of modality image instability refers to the significant variations or fluctuations in the appearance and quality of images of the same modality under different time or condition. When captured under different acquisition conditions, such as changes in lighting intensity, environmental temperature, humidity, etc., the appearance and quality of the images may be affected. These condition changes result in variations in brightness, contrast, color saturation, and other aspects of modality images, thereby affecting the

stability of the images. In addition, due to the aging, wear or inaccurate calibration of the equipment, the image quality changes, including increased noise, reduced resolution, image distortion, image overexposure and other problems, which affect the stability of the image. How to eliminate the instability of modality images in the image fusion stage of multi-modal object detection is a problem that requires further research.

5.3 Explore Fusion Based on Imaging Principle

The imaging principles of the infrared and visible light spectra differ, leading to different characteristics and information presented in their respective images. The infrared spectrum is sensitive to heat and can perceive the thermal radiation of objects. Therefore, infrared images can provide thermal information about targets, particularly in nighttime or low-light conditions, and have certain advantages in detecting thermal stealth targets. On the other hand, the visible light spectrum captures images using visible light and can provide detailed information about the shape, color, texture, and other visual characteristics of targets. Analyzing the principles of infrared and visible imaging and mathematically modeling them, along with integrating them into the image fusion stage of multi-modal object detection tasks, may be helpful in enhancing the fusion performance of images and further improving the performance of multi-modal object detection models.

6 Conclusion

This paper categorizes existing RGB-Infrared object detection models based on their fusion methods, namely pixel-level fusion, feature-level fusion, and decision-level fusion. Several models from each category are discussed in detail. Furthermore, this paper lists 5 multimodal datasets that can be used for training RGB-Infrared object detection models, providing information on the release date, image quantity, image resolution, and annotation availability. Additionally, a comparison of RGB-Infrared object detection models based on different fusion methods is conducted using the KAIST dataset. Lastly, the challenges present in the current RGB-Infrared object detection field are discussed, and future research directions are proposed.

Despite significant research progress in RGB-Infrared object detection models in recent years, the current technology still falls short of meeting practical application requirements, thus necessitating further investigation. This paper aims to provide readers with an understanding of the current research status in this field and encourage further research efforts.

Acknowledgement. This work is funded in part by the Youth Project of the National Natural Science Foundation of China under Grant 62201237, in part by the Yunnan Fundamental Research Projects under Grant 202101BE070001-008 and 202301AV070003, in part by the Youth Project of the Xingdian Talent Support Plan of Yunnan Province under Grant KKRD202203068, in part by the Major Science and Technology Projects in Yunnan Province under Grant 202202AD080013 and 202302AG050009.

References

1. Zhang, G., Luo, Z., Yu, Y., Cui, K., Lu, S.: Accelerating DETR convergence via semantic-aligned matching. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 949–958 (2022)
2. Liu, J., et al.: Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5802–5811 (2022)
3. Wang, Q., Chi, Y., Shen, T., Song, J., Zhang, Z., Zhu, Y.: Improving RGB-infrared object detection by reducing cross-modality redundancy. Remote Sens. **14**(9), 1–15 (2022)
4. Hwang, S., Park, J., Kim, N., Choi, Y., Kweon, I.S.: Multispectral pedestrian detection: Benchmark dataset and baseline. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1037–1045 (2015)
5. González, A., et al.: Pedestrian detection at day/night time with visible and FIR cameras: a comparison. Sensors **16**(6), 1–11 (2016)
6. Wagner, J., Fischer, V., Herman, M., Behnke, S., et al.: Multispectral pedestrian detection using deep fusion convolutional neural networks. In: Proceedings of European Symposium on Artificial Neural Networks (ESANN), pp. 509–514 (2016)
7. French, G., Finlayson, G., Mackiewicz, M.: Multi-spectral pedestrian detection via image fusion and deep neural networks. J. Imaging Sci. Technol. **62**, 176–181 (2018)
8. Vandersteegen, M., Van Beeck, K., Goedemé, T.: Real-time multispectral pedestrian detection with a single-pass deep neural network. In: Proceedings of Image Analysis and Recognition: 15th International Conference (ICIAR), pp. 419–426 (2018)
9. Yang, X., Qian, Y., Zhu, H., Wang, C., Yang, M.: BAANet: learning bi-directional adaptive attention gates for multispectral pedestrian detection. In: Proceedings of 2022 International Conference on Robotics and Automation (ICRA), pp. 2920–2926 (2022)
10. Kim, J.U., Park, S., Ro, Y.M.: Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection. IEEE Trans. Circ. Syst. Video Technol. **32**(3), 1510–1523 (2022)
11. Zhang, Y., Yu, H., He, Y., Wang, X., Yang, W.: Illumination-guided RGBT object detection with inter-and intra-modality fusion. IEEE Trans. Instrum. Meas. **72**, 1–13 (2023)
12. Zhuang, Y., Pu, Z., Hu, J., Wang, Y.: Illumination and temperature-aware multispectral networks for edge-computing-enabled pedestrian detection. IEEE Trans. Netw. Sci. Eng. **9**(3), 1282–1295 (2021)
13. Li, Q., Zhang, C., Hu, Q., Fu, H., Zhu, P.: Confidence-aware fusion using dempster-shafer theory for multispectral pedestrian detection. IEEE Trans. Multimedia **25**, 3420–3431 (2022)
14. Hu, Z., Jing, Y., Wu, G.: Decision-level fusion detection method of visible and infrared images under low light conditions. EURASIP J. Adv. Signal Process. **2023**(1), 1–13 (2023)
15. Guan, D., Cao, Y., Yang, J., Cao, Y., Yang, M.Y.: Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. Inf. Fusion **50**, 148–157 (2019)
16. Li, C., Song, D., Tong, R., Tang, M.: Illumination-aware faster R-CNN for robust multispectral pedestrian detection. Pattern Recogn. **85**, 161–171 (2019)
17. Zhang, H., Fromont, E., Lefevre, S., Avignon, B.: Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In: Proceedings of 2020 IEEE International Conference on Image Processing (ICIP), pp. 276–280 (2020)
18. Group, F.A.: FREE Teledyne FLIR thermal dataset for algorithm training (2018). https://www.flir.com/oem/adas/adas-dataset-form/
19. Jia, X., Zhu, C., Li, M., Tang, W., Zhou, W.: LLVIP: a visible-infrared paired dataset for low-light vision. In: Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3496–3504 (2021)