This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2024.3373816

Multi-modal Object Detection of UAV Remote Sensing Based on Joint Representation Optimization and Specific Information Enhancement

Jinpeng Wang, Congan Xu, Chunhui Zhao, Long Gao, Junfeng Wu, Yiming Yan, Shou Feng, Nan Su *Member, IEEE*

***Abstract*—With the development of Earth observation technology, it becomes easier and easier to acquire multi-modal image data at the same time. To improve the performance of multi-modal remote sensing detection algorithm, a new fusion feature optimization detection network (FFODNet) is proposed. The method is designed to solve the problem of performance degradation caused by the unreliability of single modal data in multi-modal remote sensing data. The key to obtain high quality fusion features from multi-modal data with interference is to suppress single modal redundant features and fully integrate multi-modal features. The proposed method mainly includes two improvements. Firstly, a novel joint expression optimization module (JEOM) is designed to enhance the target features and suppress the redundant and interference features that affect the fusion effect. Additionally, we propose a novel specific information enhancement module (SIEM) to further enhance the discriminative feature information of targets within each modal image. Experiments on DroneVehicle dataset show that our proposed method is state-of-the-art on this dataset.**

***Index Terms*—Multi-modal object detection, joint expression optimization module, specific information enhancement module.**

I. Introduction

THE object detection technology of Earth observation data is widely used in military and civilian fields such as intrusion warning, aerospace and so on [1], [2]. The optical image has the texture and detail information of the target, and the infrared image can provide the temperature information of the target. These two types of target information are complementary. Currently, how to make full use of multi-modal data has gradually become a new research hotspot [3]. However, in the case that there may be modal interference in multi-modal remote sensing data, how to obtain high-quality fusion features and give full play to the complementary advantages of multi-modal information is a major challenge for fusion detection technology.

This work is supported by the National Natural Science Foundation of China: No. 62271159, No. 62071136, No. 62002083, No. 61971153, No. 62371153; Excellent Youth Foundation of Heilongjiang Province of China: YQ2022F002; Fundamental Research Funds for the Central Universities: 3072022QBZ0805

Jinpeng Wang is with the College of Information and Communication Engineering, Harbin Engineering University, China, 150001, and also with the Naval Aeronautical University China, 264000.(e-mail: wangjinpeng9521@hrbeu.edu.cn)

Chunhui Zhao, Nan Su, Yiming Yan and Shou Feng are with the College of Information and Communication Engineering, Harbin Engineering University, China, 150001, and also with Key Laboratory of Advanced Marine Communication and Information Technology, Ministry of Industry and Information Technology, Harbin 150001, China. (e-mail: zhaochunhui@hrbeu.edu.cn; yanyiming@hrbeu.edu.cn; fengshou@hrbeu.edu.cn; sunan08@hrbeu.edu.cn;)

Congan Xu, Long Gao and Junfeng Wu are with the Naval Aeronautical University China, 264000.(e-mail: xcatougao@163.com; 1017730430@qq.com; patrickwu0609@163.com;).

(Corresponding author: Congan Xu and Nan Su.)

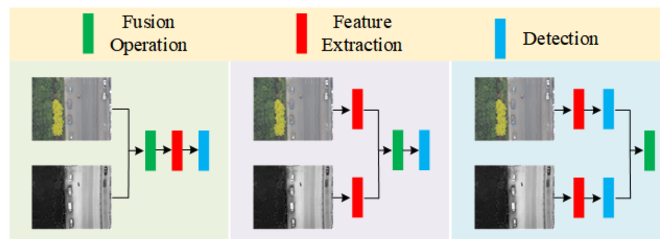


Fig. 1. Diagram of three types of fusion scheme. (a) Detection by fusion image; (b) Detection by fusion feature; (c) Fusion both detection result.

Optical remote sensing images have the advantages of easy access and high resolution. Many researchers use optical remote sensing images for object detection [4], [5]. However, in some challenging visual scenes, such as low illumination, smoke interference, and so on, relying solely on optical images for detection often fails. Infrared remote sensing images can obtain temperature information of the target without relying on visual factors in the environment. Some researchers have focused on using infrared images for object detection [6]. Due to the low resolution of infrared image, it is difficult to detect difficult targets such as low contrast, small scale and lack of texture. In the face of some highly difficult suspected targets, even the human eye is difficult to judge. If the temperature information of the infrared image and the details and colors of the optical image can be used at the same time, the detection performance can be greatly improved. Therefore, it is worth exploring how to solve the inherent limitations of single-modal data by using multi-modal complementary information to improve the detection performance [7], [8].

At present, deep learning technology is widely used in various fields [9], and it is also the focus of research in the field of multi-modal target detection. Figure 1 shows different fusion strategies in multi-modal object detection algorithms in detail, which are image-level fusion, feature-level fusion and decision-level fusion. In this figure, the red part is responsible for feature extraction, the green part represents the fusion step, and the blue part points to the object detection. Many studies have shown that implementing multi-modal feature fusion in

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2024.3373816

the middle layer of the network can usually obtain better multi-modal detection results [10]. Nowadays, multi-modal object detection methods based on feature fusion have been widely concerned and become the mainstream trend.

To use optical and infrared remote sensing images for object detection, some researchers carry out direct weighted summation of each modal branching feature. Li et al. introduce illumination information to guide the fusion of multi-modal image features [11]. They use a simple CNN and prediction head to evaluate the illumination of RS optical images, and according to the evaluation results, the importance of optical image branch features is obtained, and finally the weighted summation of each mode feature is carried out to obtain the fusion feature. Guan et al. went a step further by using a deeper hierarchy to evaluate the illuminance information [12], and then used the evaluation results to guide multi-modal feature fusion. However, the branch learning efficiency of the modal weight calculation of this fusion method is low, which affects the detection performance.

In order to obtain the fusion features adaptively, Tu et al. extracted the fusion feature information by layer by layer joining and convolution of multi-modal features [13], which improved the overall efficiency of the algorithm. Zhang et al. went further, using a more complex concatenation convolutional structure to obtain fusion features of multiple subspaces simultaneously [14], and finally grouping these features along the channel dimension to obtain fusion features for prediction. This approach directly combines the features of the two branches and ignores the possible interference in the modal information.

Recently, attention structures have been widely used in various architectures because of their excellent performance. Meng uses a residual attention structure to perform self-attention operations on convolution-acquired fusion features to highlight their useful components [15]. Zhao et al. uses Transformer architecture to perform self-attention operations on multi-modal features and perform attention weighting in multiple subspaces [16]. Wang et al. applied the attention structure to feature fusion networks combined with reliability weighting operations for high-level semantic information [17]. This method combines the high efficiency of multi-modal detection based on feature fusion with the robustness based on reliability weighting, and has achieved great success on a remote sensing dataset. However, in the above methods, the attention structure is only applied to the feature fusion method, while the backbone network still uses the general structure for feature extraction.

In summary, multi-modal object detection still faces many challenges due to the significant modal differences between different modal data. The simple weighted fusion method struggles to fully aggregate the multi-modal feature information. The fusion method based on concatenation convolution fails to consider how to suppress feature information that hinders fusion. Although the fusion methods based on attention structure design have shown excellent performance, they neglect to improve the single-modal feature extraction ability of the backbone network, ultimately reducing the efficiency of multi-modal feature fusion. Additionally, optical images can introduce interference in the feature fusion process, especially under low illumination conditions. The discriminative information is the information that can distinguish the target from the background, such as the temperature difference between the target and the background in the infrared image, and the color difference in the optical image. The simple feature fusion methods may inadvertently introduce interference information and reduce the overall detection performance.

The main contributions of this paper are as follows:

1. To address the aforementioned challenges, this paper proposes a novel two-branch multi-modal detection Fusion feature Optimization detection network (FFODNet). The FFODNet aims to adaptively fuse target feature information from multi-modal remote sensing images and achieve high-performance detection. Specifically, the method consists of two key improvements: the backbone network and the feature fusion module.

2. To fully integrate multi-modal features and suppress interference information that is unfavorable to fusion, we propose a joint expression optimization module (JEOM) based on cross-concern. The JEOM is designed to adaptively extract high-quality multi-modal joint expressions of objects of interest from remote sensing data with uncertain primary and secondary states.

3. To enhance the discriminative feature information of the target in the single-modal image, we propose a new specific information enhancement module (SIEM) in the two-branch backbone network. The SIEM is designed to suppress irrelevant background feature information and further improve the efficiency of the subsequent feature fusion operation.

The rest of this article is structured as follows: in Section II, we describe the network structure and methods in detail. Section III gives the details of our work and experimental results and related comparison to verify the effectiveness of our method. Finally, we summarize the research content in the Section IV.

II. Proposed Method

The overall architecture of the proposed detection method is illustrated in Figure 2. Since the infrared image and optical image have a similar data format, we utilize an isomorphic backbone network to extract features from the multi-modal images. To ensure that the image features of each modality have the same dimension within the network, we expand the single-channel infrared image to three channels by duplicating the same value across all channels. The feature extraction process begins with a double-branch structure, which extracts features from the multi-modal images. Subsequently, the joint expression optimization module (JEOM) utilizes these features to suppress information that is not conducive to fusion, thereby extracting high-quality joint feature representations. Simultaneously, the specific information enhancement module (SIEM) is employed to enhance the extracted features, thereby improving the discriminative features of each individual modality and further enhancing overall performance. Finally, the fused features are passed to the detection head, where the detection results are obtained. This detection head leverages the fused features to identify and localize the target objects in the scene.

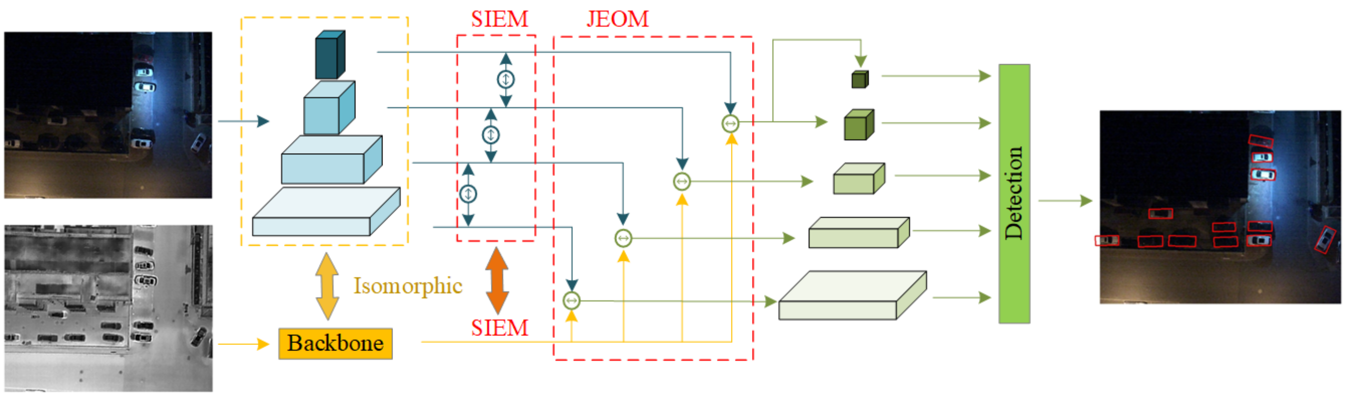


Fig. 2. The proposed multi-modal object detection method mainly includes backbone, SIEM, JEOM and detection.

The proposed architecture combines multi-modal feature fusion and single-modal feature enhancement, allowing the network to effectively capture useful complementary information in multi-modal remote sensing images and improve the performance of detection tasks. Additionally, the isomorphic backbone network enables the extraction of consistent and compatible features from different modalities, facilitating the fusion process and enhancing the overall performance of the detection method.

The specifics of the JEOM and SIEM are elaborated upon in Section II-A and Section II-B, respectively.

A. Joint Expression Optimization Module

To obtain high-quality multi-modal fusion features, we propose the JEOM based on cross attention to address possible interference in remote sensing data. This module incorporates both single-modal features and multi-modal joint features to perform attention operations. The objective of these operations is to enhance useful information while suppressing redundant information that is not beneficial for fusion. Networks generally exhibit better performance when they have access to more useful information. By incorporating cross attention mechanisms within the JEOM, our proposed method effectively focuses on important features and enhances their representation in the fusion process. This allows the network to leverage the most relevant and discriminative information from both single-modal and multi-modal features, thereby leading to improved overall performance.

The query tensor is used to search for and enhance useful features in each single mode. If the query tensor can search for the target feature more accurately, then it can also be regarded as learning a lot of useful information. To further improve the fusion efficiency, an additional step is designed during the fusion process. This involves adding the query vector, which is calculated from each single-modal feature, to the fusion feature. The structure of this enhanced fusion feature is depicted in Figure 3.

As shown in Figure 3, the optical and infrared features are input to the fusion module and mapped as query tensors respectively. These query tensors are then combined with the simple fusion multi-modal features for attention operations.

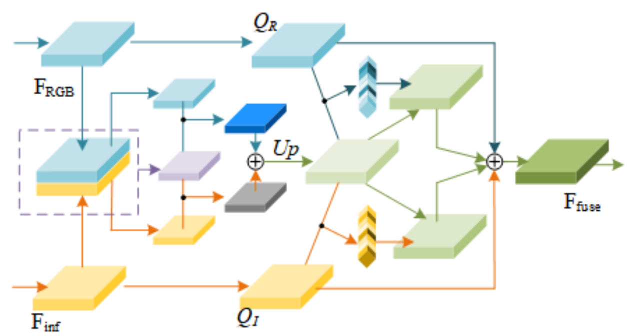


Fig. 3. Schematic diagram of joint expression optimization module (JEOM).

Finally, the query tensor and the enhanced fusion feature are added together to obtain the final fusion feature. The entire process leverages the attention mechanism to suppress redundant information, thereby obtaining a high-quality fusion feature expression.

By incorporating this enhanced fusion feature into the network architecture, aiming to further emphasize the useful information contained within the single modal features, thereby increasing the proportion of useful information in the final fusion feature.

In addition, an attention-enhancing structure similar to the Transformer strategy is designed, which has a strong ability to acquire high-quality fusion features [18]. In this structure, in order to carry out adaptive information fusion, we use the form of convolution instead of tensor product operation. Firstly, the module fuses the multi-modal input features, obtains the preliminary fusion features, and regards them as key tensors and value tensors. Secondly, the query tensor is calculated using each modal feature. Then, two query tensors and key tensors are used to calculate the weight vector respectively, and the value tensor is weighted by the weight vector. Finally, the enhanced fusion features are added to the query tensor to obtain the final fusion features. Finally, the enhanced fusion feature is added to the query tensor to obtain the final fusion feature. The formula of the overall calculation process is expressed as follows:

$$JEOM(F_i^{RGB}, F_i^{Inf}) = \tilde{F_i^{Inf}} + \tilde{F_i^{RGB}} + Q_I + Q_R \quad (1)$$

where $F_i^{\tilde{Inf}}$ and $F_i^{\tilde{RGB}}$ refer to the enhanced feature tensor respectively. Q_R and Q_I refer to query tensors calculated from RGB image features and infrared image features, respectively.

$$Q_R, Q_I = Conv_{1\times1}(F_i^{RGB}), Conv_{1\times1}(F_i^{Inf}) \quad (2)$$

The $Conv_{1\times1}$ represents a convolution operation with kernel size 1 that does not change the dimension. Take the feature calculation of optical image as an example, the formula is as follows:

$$F_i^{\tilde{RGB}} = Tr(Q = Q_R, K = F_i^f, V = F_i^f) \quad (3)$$

The Tr is a cross-attention enhancement operation similar to the Transformer strategy. It uses the query tensor and the key tensor to calculate the weight vector, and weights the value tensor as follows:

$$Tr(Q, K, V) = W(Q, K) \cdot V \quad (4)$$

The dot multiplication in the formula refers to the multiplication of values along the channel dimension after broadcasting, so as to achieve the purpose of feature selection.

$$W(Q, K) = Pool_{average}[CBL_{1\times1}(Q, K)]_{h,w} \quad (5)$$

$$CBL_1(X, Y) = L\_ReLu\{Bn\{Conv_1[Cat(X, Y)c]\}\} \quad (6)$$

The above formula represents the use of weight vector to enhance the channel dimension of the feature graph. Where $Pool_{average}$ refers to the global averaging pooling operation of feature tensors along the width and height directions. The $CBL_{1\times1}$ refers to the concatenation convolution operation. The Bn and L_ReLu are the batch normalized operation and the activation operation using the Leaky ReLu function, respectively. In the above equation, the F_i^f refers to the preliminary fusion feature, and the calculation formula of it is as follows:

$$F_i^f = Upsample(\tilde{F_i^r} + \tilde{F_i^i}) \quad (7)$$

The primary fusion features are obtained by adding and applying operations on the calculated $\tilde{F_i^r}$ and $\tilde{F_i^i}$.

$$\tilde{F_i^r} = Conv_{1\times1}(Cat(F_i^r, F_i^{add})) \quad (8)$$

$$\tilde{F_i^i} = Conv_{1\times1}(Cat(F_i^i, F_i^{add})) \quad (9)$$

To fully capture the discriminative information from each modality, we concatenate the optical feature and infrared feature and perform convolutions with the fusion feature separately. This process enables the network to focus not only on the combined features but also on individual modal features, thus providing essential information for subsequent fusion operations. the resulting fusion feature map is denoted that primarily emphasizes the optical modal features as $\tilde{F_i^r}$. The F_i^{add} in the formula refers to the multi-modal features of the initial fusion. This step enhances the discriminative capability of the network and contributes to improved multi-modal object detection performance.

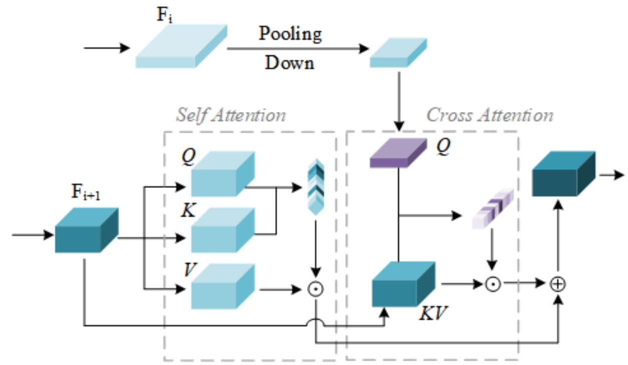


Fig. 4. Schematic diagram of specific information enhancement module (SIEM).

$$F_i^{add} = Conv_{3\times3}^{s=2}(F_i^{RGB} + F_i^{Inf}) \quad (10)$$

$$F_i^r, F_i^i = Conv_{3\times3}^{s=2}(F_i^{RGB}), Conv_{3\times3}^{s=2}(F_i^{Inf}) \quad (11)$$

The $Conv_{3\times3}^{s=2}$ in the formula represents a convolution operation with kernel size 3 and step size 2. In order to obtain the context information of different modal features, we carry out a downsampling operation on the features involved in the calculation when calculating the initial fusion features, so as to obtain useful information conducive to fusion in a larger area.

B. Specific Information Enhancement Module

To enhance the feature extraction capability of single-modal remote sensing images, we propose a self-attention SIEM module to further improve the detection performance. This module leverages multi-scale feature information to enhance the discriminative features of the target, thereby improving the network's attention to target information during the feature fusion stage. We adopt the attention-enhancing structure proposed in this article within the SIEM. Given that the input features consist of both deep and shallow features that exhibit strong correlations, the operation methods of the query tensor, key tensor, and value tensor within the attention network in this section differ from those introduced in the previous section. These modifications are necessary to effectively capture and enhance the discriminative information within the multi-scale feature representation. By incorporating the self-attention SIEM module, we enable the network to dynamically emphasize target-related information and enhance the discriminative power of the fused features.

The structure diagram of SIEM is shown in Figure 4. SIEM utilizes shallower features to enhance attention towards deeper features. In many feature extraction networks, downsampling operations are performed on feature maps to reduce computation. However, this downsampling process can result in the loss of certain low-level spatial location information. SIEM addresses this issue by enhancing the discriminative features of the target and reducing the loss of spatial position information caused by downsampling during feature extraction.

In SIEM, the input shallow feature is scaled, and the module performs separate mapping operations on the deep feature

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2024.3373816

and the scaled shallow feature. Specifically, key features and query features are mapped, and their interaction generates an attention vector. This attention vector is then utilized to enhance the discriminative feature information within the deep feature.

By incorporating SIEM into the network structure, we leverage richer spatial information from the shallower features to enhance the deeper features. Simultaneously, the deeper features undergo self-attentional operations to enhance useful target feature information. The useful information component in the feature is enhanced to improve the efficiency of the subsequent feature fusion network.

The overall calculation process is as follows:

$$SIEM(F_{i-1}, F_i) = Att_{self}(F_i) + Att_{cro}(F_{i-1}, F_i) \quad (12)$$

The F_i and F_{i+1} represent shallow features and deep features in the process of feature extraction. The output of this module is the sum of the arithmetic result of the deep feature self-attention enhancement and the weighted feature of the shallow feature. The formula for self-attention enhancement is as follows:

$$Att_{self}(F_i) = Tr(Q_s(F_i), K_s(F_i), F_i) \quad (13)$$

$$Q_s(F_i), K_s(F_i) = Conv_{1\times1}(F_i), Conv_{1\times1}(F_i) \quad (14)$$

In the cross-attention operation, deep and shallow features need to be aligned in wide and high dimensions. In order to retain more low-level spatial information, we only use one convolutional layer for dimensional alignment. Although the method also downsamples shallow features, compared with the backbone network, the single-layer convolution operation can retain more spatial information rather than extract more semantic information. In addition, the single-layer convolution structure can also establish additional residual paths and improve the training efficiency of the method. The cross-attention formula is as follows:

$$Att_{cro}(F_{i-1}, F_i) = Tr(Q_c(F_{i-1}), K_c(F_i), F_i) \quad (15)$$

$$Q_c(F_{i-1}), K_c(F_i) = Conv_{3\times3}^{s=2}(F_{i-1}), Conv_{1\times1}(F_i) \quad (16)$$

The Q_c and K_c in the formula refer to the computational structure of the query tensor and the bond tensor in this part. The operation process of the function Tr in the formula is the same as the formula (4) in Section II-A.

In the overarching process, we input the multi-modal image data into the network and initiate distinct feature extraction procedures using a specialized isomorphic backbone network tailored for each modality. This step enables us to capture and emphasize unique characteristics inherent to each type of data. Following this initial extraction, we propose the SIEM to dynamically amplify the discriminative information present within the features of each individual modal. This enhancement process occurs independently within dedicated branches for each modal.

Building upon this enhancement, we activate the JEOM, which orchestrates the fusion of multi-modal features at the same hierarchical level. The objective here is to ensure a harmonious integration of information across modalities, promoting a synergistic representation. Within this fusion process, careful attention is given to utilizing each modal feature map in order to selectively suppress redundant information embedded within the initial fusion feature map. This approach aims to distill and preserve only the most pertinent details, generating fusion features of elevated quality.

In culmination of this sophisticated process, we deploy resulting high-quality fusion features for critical object detection tasks. Our comprehensive approach encompasses distinct feature extraction, individual modal enhancement, multi-modal fusion, and information refinement, all working collectively to ensure the robust and effective performance of our network in discerning and identifying targets within multi-modal image data. By taking advantage of the different strengths of each mode, our network extracts target discrimination features from multi-modal images that capture both multi-modal shared information and single-modal specific information. The individual modal enhancement module refines the features of each modality, boosting their discriminative power and facilitating more accurate object detection. The multi-modal fusion process effectively integrates the enhanced features from different modalities, enabling the network to exploit the complementary information and achieve a more comprehensive understanding of the scene.

III. EXPERIMENT

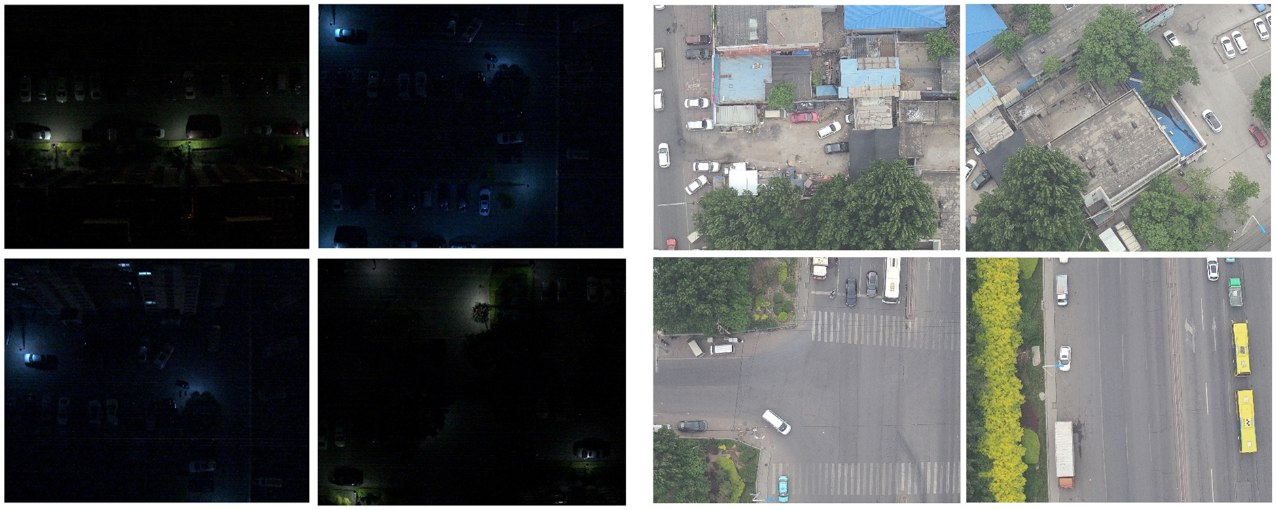
A. Expereimental Datasets

The DroneVehicle dataset consists of 19,459 pairs of RGB-infrared images, classified as vehicles, captured by camera-equipped drones [19]. Regional scenes are divided into urban roads, residential areas and highways. The lighting conditions were night and day. We used 17,990 RGB-infrared image pairs for training, 1,469 pairs for validation. The overall dataset contains the following five categories of objectives: car, freight car, truck, bus and van. Some images have the problem of low contrast or low illumination, which will cause the network to be interfered by certain modal data in the feature fusion stage, which requires the feature balancing performance of the network. According to the interference degree of modal data, the data set is divided into the following two parts: the weak interference subset and the strong interference subset. The subsets are shown as Figure 5. It is important to note that we have only divided the test set.

TABLE I
VOLUME DIAGRAM TABLE FOR EACH DATASET.

Dataset	classes	train set	test set	tote
DroneVehicle	5	17,990	1,469	19,459
FLIR-aligned	3	4,129	1,013	5,142

To ensure the consistency of the image scale between the two modalities in FLIR, we conducted experiments on the 'aligned' version [20]. The 'aligned' FLIR contains 5,142 RGB-infrared image pairs, of which we used 4,129 pairs for training and 1,013 pairs for testing. It covers different urban



(a) (b)

Fig. 5. The schematic samples of the weak interference subset and the strong interference subset are shown in the figure. In (a) is the schematic data of the strong interference subset, and in (b) is the schematic data of the weak interference subset.

street scenes and includes three object categories: bicycle, car, and person.

The training, testing, and overall data volumes for each dataset are shown in the table I.

B. Implementation Details

We executed all experiments using PyTorch on a machine equipped with a GeForce RTX 3090 GPU. The optimization process employed the Stochastic Gradient Descent (SGD) algorithm, with an initial learning rate set at 0.003, an attenuation weight of 0.0001, and a momentum value of 0.9. To quantitatively assess the performance of multi-modal object detection, we employed the conventional evaluation metric known as average mean precision (mAP).

C. Performance Evaluation on DroneVehicle dataset

Our base-structure is a two-branch object detection network without SIEM and JEOM in Figure 2. The baseline of the proposed method is improved by the single branch Faster-RCNN [21]. To explore a better feature fusion method, we conducted some relative comparative experiments. Hong et al. listed a number of multi-modal feature fusion methods [22]. Sharma et al. and Zhang et al used Point-wise addition and Concat-Conv respectively to carry out multi-modal feature fusion, and achieved certain results [23], [24]. We reproduced the most complex fusion method in their paper and compared it with direct addition and concatenation convolution, two simple fusion methods, their structures are shown in Figure 6.

These three fusion methods are called Point-wise addition, Concat-Conv and Cross-Concat-Conv in turn. Each of the methods in the table uses ResNet50 for feature extraction of single-modal images and only uses different structures for feature fusion. Finally, the fused features are used for object detection. By comparing the performance differences of various fusion methods, experiments show that the proposed fusion method is superior to the above fusion methods.

Table II shows that a more complex feature fusion network can obtain better multi-modal fusion features, thus improving the performance of the object detection method. To improve the learning efficiency of the network and make a more explicit performance comparison, we refer to the detection network using direct addition operation in the fusion part as the baseline.

TABLE II
THE EXPERIMENTAL PERFORMANCE OF EACH OBJECT DETECTION METHOD ON DRONEVEHICLE DATASET, AS WELL AS THE MODAL IMAGES IT USES. THE OPTIMAL DETECTION RESULTS ARE SHOWN IN BOLD.

Method	Modality	mAP
Faster R-CNN [21]	R\I	43.94\52.63%
RoITransformer [25]	R\I	47.91\59.15%
ReDet [26]	R\I	51.04\60.54%
Gliding Vertex [27]	R\I	52.48\62.89%
Point-wise addition [23]	R+I	60.82%
Concat-Conv [24]	R+I	61.63%
Cross-Concat-Conv [22]	R+I	64.24%
UA-CMDet [19]	R+I	64.01%
RISNet [17]	R+I	66.40%
AR-CNN [28]	R+I	71.58%
FFODNet(ours)	R+I	**76.93%**

And table II shows the comparison between the proposed method and the current object detection method with multi-modal feature fusion capability. RISNet and UA-CMDet are good fusion object detection algorithms at present [17], [19], both using a mixture of feature-level fusion and decision-level fusion strategies, but their performance is still inferior to our proposed method.

For better comparison, the detection structure of the (a) method in Figure 6 described above is used as our baseline. Our improved mAP improves by about 16% compared to baseline results and is significantly higher than the single-modal object detection method. Experiments show that the proposed method can extract high quality fusion feature information

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2024.3373816

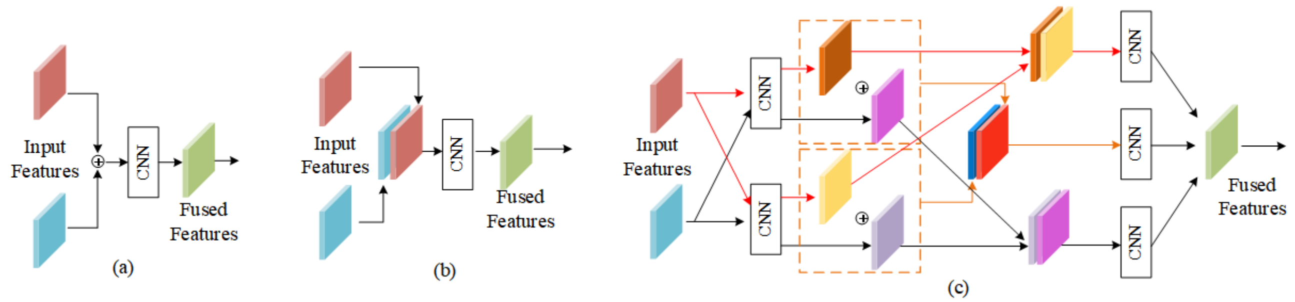


Fig. 6. Comparison of three fusion structure diagram. (a) is the direct addition of the feature map; (b) Convolution after concatenation of feature graphs; (c) refers to multiple interleaved concatenation and convolution of feature graphs.

from infrared-optical image pairs, and its performance is state of the art.

The subjective detection results of each algorithm are presented in Figure 7. The first line represents the fusion method of Point-wise addition, the second line represents the fusion method of Concat-Conv, and the third line represents the fusion method of Cross-Concat-Conv. The fourth line displays the detection result of UA-CMDet, and the image is sourced from the original paper. The final line demonstrates the test results of the proposed method. The proposed method demonstrates excellent detection effectiveness, performing well on dense targets and targets with obstructed edges

The quantitative analysis of each fusion method on the two subsets is compared in the III. To evaluate the robustness of the algorithm in difficult scenarios, the test set of the original data set is divided into two test subsets according to the difficulty of the scenario. Experiments show that the proposed method is robust in low illumination environment.

TABLE III
THE EXPERIMENTAL PERFORMANCE OF EACH OBJECT DETECTION METHOD ON DRONEVEHICLE DATASET, AS WELL AS THE MODAL IMAGES IT USES. THE OPTIMAL DETECTION RESULTS ARE SHOWN IN BOLD.

Method	mAP in subset (a)	mAP in subset (b)
Point-wise addition [23]	59.8%	65.5%
Concat-Conv [24]	59.4%	72.9%
Cross-Concat-Conv [22]	64.2%	77.3%
FFODNet(ours)	**75.2%**	**83.2%**

As shown in Figure 8, the true value is shown in green on the first row of the RGB image. The second and third rows show the detection results for the baseline and the FFODNet, respectively. To enhance visual clarity, we highlight the objects with our approach above the baseline in yellow.

Visualizations of some of the detection results in the strong interference subset are depicted in Figure 9. In the first column of the presented data, the infrared data exhibits a low contrast phenomenon, while the optical image suffers from cloud interference. The feature fusion network needs to address the low illumination interference caused by optical images in the second and fourth columns of data. In the third column of data, both cloud and low illumination problems are observed in the optical images. These phenomena indicate that a certain modality of data is not always reliable in the fusion object detection task. Consequently, it is crucial for the algorithm to initially treat all modal data as equally important during feature fusion and adaptively suppress the interference introduced by modal data throughout the fusion process.

The experimental results unequivocally validate the effectiveness and feasibility of our algorithm, particularly when applied to the challenging strong interference subset. The algorithm demonstrates excellent detection performance, surpassing expectations and showcasing its potential for real-world applications. The exceptional performance of our algorithm can be attributed to its innovative fusion strategy, adaptive weighting mechanism, and end-to-end training approach. These key components enable the algorithm to effectively handle interference from different modalities, prioritize relevant information, and optimize the feature fusion process for accurate object detection. The algorithm's ability to adaptively suppress interference and exploit the complementary characteristics of multi-modal data contributes to its outstanding performance.

D. Ablation Experiment on DroneVehicle dataset

To verify the effectiveness of the module proposed in this paper, we conducted ablation experiments on JEOM and SIEM. Table IV shows the experimental results of the ablation experiments. The baseline in Table IV refers to the method Point-wise addition in Table II.

TABLE IV
OUR PROPOSED METHOD WAS COMPARED WITH BASELINE ABLATION EXPERIMENTS. THE OPTIMAL DETECTION RESULTS ARE SHOWN IN BOLD.

Method	car	freight-car	truck	bus	van	mAP
baseline	89.70%	35.90%	49.00%	88.30%	41.20%	60.82%
base+S	90.10%	52.60%	64.20%	88.20%	54.50%	69.92%
base+J	90.30%	62.30%	71.70%	89.40%	60.70%	74.88%
base+J+S	**90.40%**	**68.40%**	**72.60%**	89.20%	**64.10%**	**76.93%**

The J and S in Table IV refer to the JEOM and SIEM in Section II of this article. The fusion module is compared with the simple fusion method. Furthermore, SIEM is used to improve the ability of the network to extract the discriminant features of the target, so as to improve the efficiency of feature fusion. In the proposed method, the detection efficiency is improved through the synergistic effect of JEOM and SIEM. Experiments show that JEOM can optimize the detection performance by improving the efficiency of feature fusion. On this basis, SIEM can further improve the overall network performance. On the basis of optimizing the feature fusion

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2024.3373816

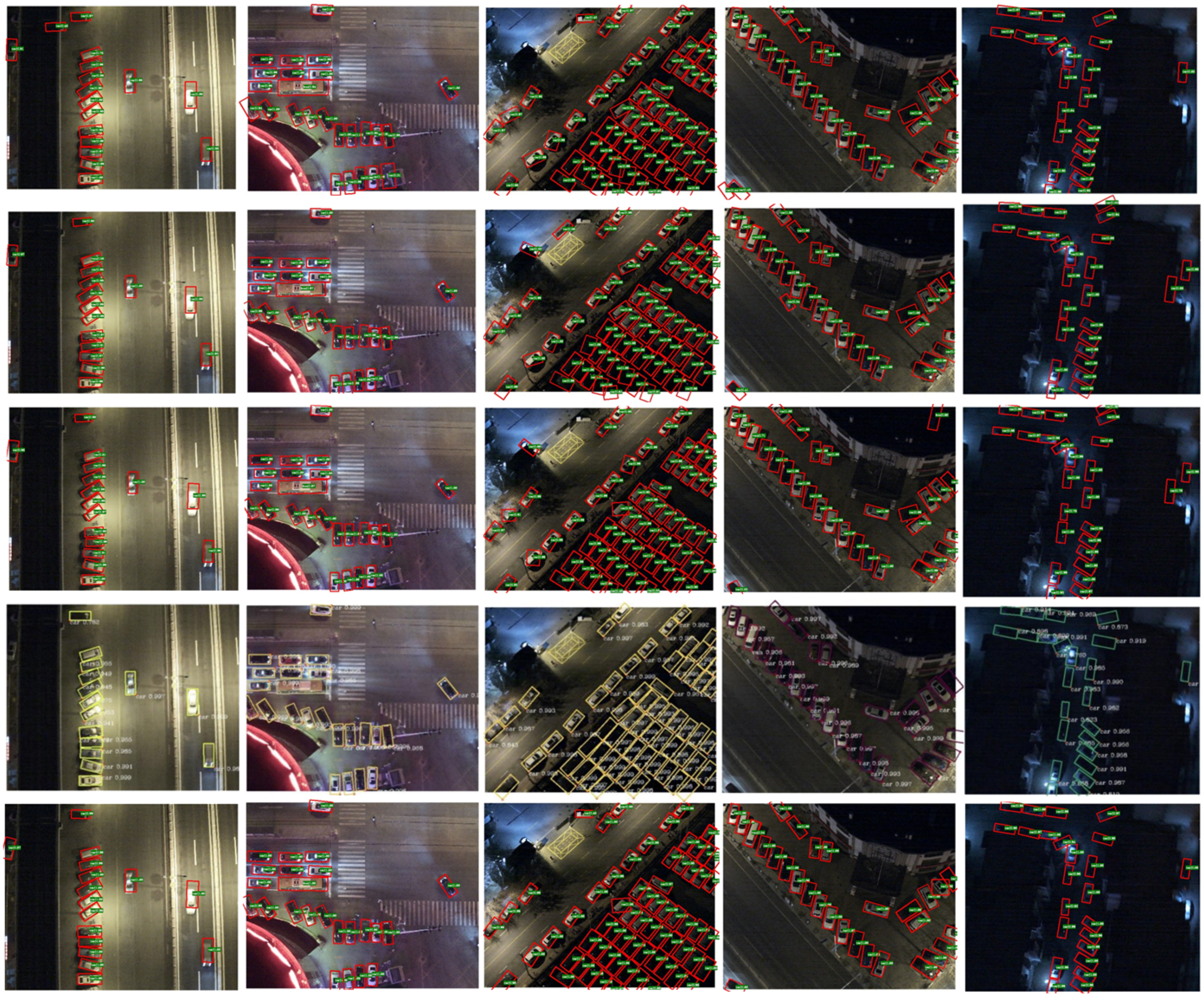


Fig. 7. Visualization of detection results of each algorithm. From top to bottom are the test results of Point-wise addition method, Concat-Conv, Cross-Concat-Conv, UA-CMDet and FFODNet.

structure, it is meaningful to enhance the ability of each feature extraction branch.

E. Performance Evaluation on FLIR dataset

To further validate the effectiveness of our proposed method, we conducted comparison experiments with other state-of-the-art methods on the FLIR-aligned dataset.

TABLE V
THE EXPERIMENTAL PERFORMANCE OF EACH OBJECT DETECTION METHOD ON FLIR DATASET, AS WELL AS THE MODAL IMAGES IT USES

Method	Modality	mAP
Faster R-CNN [21]	R\I	63.60\75.30%
HalfwayFusion [29]	R+I	71.17%
DALFusion [30]	R+I	72.11%
CFR [29]	R+I	72.39%
GAFF [31]	R+I	73.80%
YOLO-MS [32]	R+I	75.20%
MFF-YOLOv5 [15]	R+I	78.20%
UA-CMDet [19]	R+I	78.60%
FFODNet(ours)	R+I	78.30%

Table V presents the results of several advanced object detection methods that possess multi-modal fusion capabilities. As observed from Table V, our approach outperforms the other methods, establishing itself as the leading method for object detection on the FLIR dataset. These results further demonstrate the superior performance and effectiveness of our proposed method in multi-modal object detection tasks.

As observed in Table 3, using only infrared modal images can achieve a higher degree of precision, mainly because the infrared images in the FLIR dataset provide a better view compared to the optical images. Inadequate fusion methods may introduce interference information that hinders fusion, ultimately reducing the detection performance of the network. Notably, the YOLO-MS is recently advanced multi-modal fusion object detection method. However, in our experiments, we have achieved higher performance compared to the method. Furthermore, while our approach performs equally well as UA-CMDet on the FLIR dataset, it outperforms UA-CMDet on the drone dataset. These results highlight the superior performance and effectiveness of our approach in both FLIR and drone

Fig. 8. Subjective diagram of test results. The rows from top to bottom are labeled optical images, baseline detection results, and the results of the FFODNet.

datasets.

IV. Conclusion

In this paper, a novel multi-modal detection fusion feature optimization detection network (FFODNet) is proposed, which adaptively fuses the target feature information of multi-modal remote sensing images to achieve high performance detection. It includes the improvement of backbone network and fusion module. In order to obtain high quality fusion features by enhancing object-specific features and suppressing redundant information that may hinder fusion, a new JEOM is proposed. Based on this, a new SIEM is designed to suppress irrelevant background feature information and further improve the efficiency of subsequent feature fusion operations. Experimental results show that our proposed method outperforms existing state-of-the-art methods on the DroneVehicle dataset.

References

[1] W. Lu, C. Lan, C. Niu, W. Liu, L. Lyu, Q. Shi, and S. Wang, "A cnn-transformer hybrid model based on cswin transformer for uav image object detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 1211–1231, 2023.

[2] Y. Han, W. Meng, and W. Tang, "Capsule-inferenced object detection for remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 5260–5270, 2023.

[3] J. Xue, D. He, M. Liu, and Q. Shi, "Dual network structure with interweaved global-local feature hierarchy for transformer-based object detection in remote sensing image," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 6856–6866, 2022.

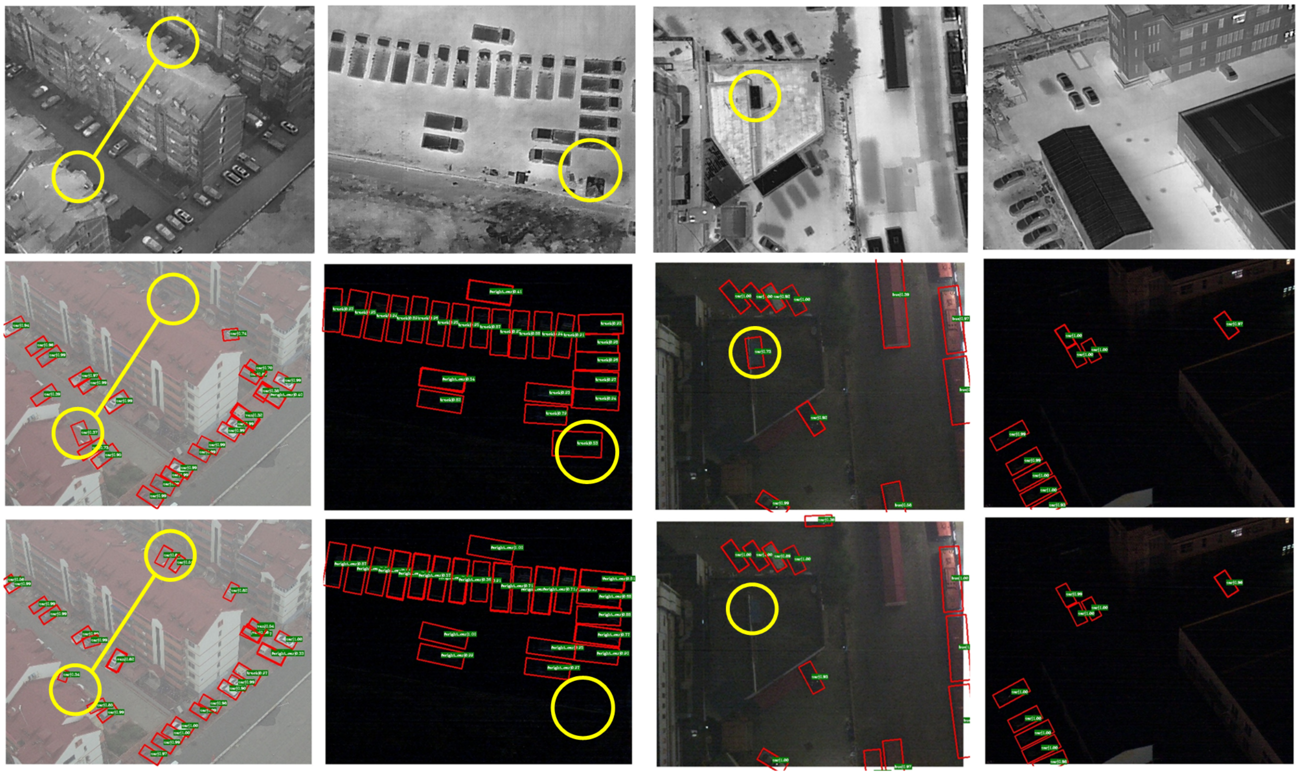


Fig. 9. Figure of data sample detection results with modal data interference. The first is the infrared image data, the second is the detection result of the baseline, and the third line is the detection result of the proposed method.

[4] S. Tian, L. Kang, X. Xing, J. Tian, C. Fan, and Y. Zhang, "A relation-augmented embedded graph attention network for remote sensing object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2022.

[5] Y. Tang, S. Feng, C. Zhao, Y. Fan, Q. Shi, W. Li, and R. Tao, "An object fine-grained change detection method based on frequency decoupling interaction for high-resolution remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024.

[6] D. Zhou and X. Wang, "Robust infrared small target detection using a novel four-leaf model," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 1462–1469, 2024.

[7] Y. Lin, Y. Tu, Z. Dou, L. Chen, and S. Mao, "Contour stella image and deep learning for signal recognition in the physical layer," *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 1, pp. 34–46, 2021.

[8] Q. Wang, M. Wang, Z. Zhang, J. Song, K. Zeng, T. Shen, and Y. Gu, "Multispectral point cloud superpoint segmentation," *SCIENCE CHINA Technological Sciences*, pp. –, 2023.

[9] Y. Tu, Y. Lin, C. Hou, and S. Mao, "Complex-valued networks for automatic modulation classification," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 9, pp. 10 085–10 089, 2020.

[10] Q. Wang, C. Yin, H. Song, T. Shen, and Y. Gu, "Utfnet: Uncertainty-guided trustworthy fusion network for rgb-thermal semantic segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[11] 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.

[12] 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.

[13] Z. Tu, Y. Ma, Z. Li, C. Li, J. Xu, and Y. Liu, "Rgbt salient object detection: A large-scale dataset and benchmark," *IEEE Transactions on Multimedia*, vol. 25, pp. 4163–4176, 2023.

[14] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, and Q. Du, "Superyolo: Super resolution assisted object detection in multimodal remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[15] S. Meng and Y. Liu, "Multimodal feature fusion yolov5 for rgb-t object detection," in *2022 China Automation Congress (CAC)*, 2022, pp. 2333–2338.

[16] C. Zhao, H. Liu, N. Su, and Y. Yan, "Tftn: A transformer-based fusion tracking framework of hyperspectral and rgb," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[17] Q. Wang, Y. Chi, T. Shen, J. Song, Z. Zhang, and Y. Zhu, "Improving rgb-infrared object detection by reducing cross-modality redundancy," *Remote Sensing*, vol. 14, no. 9, 2022.

[18] C. Zhao, H. Liu, N. Su, C. Xu, Y. Yan, and S. Feng, "Tmtnet: A transformer-based multimodality information transfer network for hyperspectral object tracking," *Remote Sensing*, vol. 15, no. 4, 2023.

[19] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6700–6713, 2022.

[20] H. Zhang, É. Fromont, S. Lefèvre, and B. Avignon, "Multispectral fusion for object detection with cyclic fuse-and-refine blocks," *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 276–280, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221970539

[21] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, 2015.

[22] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4340–4354, 2021.

[23] M. Sharma, M. Dhanaraj, S. Karnam, D. G. Chachlakis, R. Ptucha, P. P. Markopoulos, and E. Saber, "Yolors: Object detection in multimodal remote sensing imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 1497–1508, 2021.

[24] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, and Q. Du, "Superyolo: Super resolution assisted object detection in multimodal remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[25] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning roi transformer for oriented object detection in aerial images," in *2019*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2844–2853.

[26] J. Han, J. Ding, N. Xue, and G. Xia, "Redet: A rotation-equivariant detector for aerial object detection," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2785–2794, 2021.

[27] Y. Xu, M. Fu, Q. Wang, Y. Wang, K. Chen, G.-S. Xia, and X. Bai, "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 4, pp. 1452–1459, 2021.

[28] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, and Z. Liu, "Weakly aligned cross-modal learning for multispectral pedestrian detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 5126–5136.

[29] H. Zhang, E. Fromont, S. Lefevre, and B. Avignon, "Multispectral Fusion for Object Detection with Cyclic Fuse-and-Refine Blocks," *arXiv e-prints*, p. arXiv:2009.12664, Sep. 2020.

[30] H. Zhang, E. Fromont, S. Lefevre, and B. Avignon, "Deep active learning from multispectral data through cross-modality prediction inconsistency," in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 449–453.

[31] H. ZHANG, É. Fromont, S. Lefèvre, B. Avignon, and U. de Rennes, "Guided attentive feature fusion for multispectral pedestrian detection," *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 72–80, 2021.

[32] Y. Xie, L. Zhang, X. Yu, and W. Xie, "Yolo-ms: Multispectral object detection via feature interaction and self-attention guided fusion," *IEEE Transactions on Cognitive and Developmental Systems*, pp. 1–1, 2023.