Research on Multi-modal Object Detection Methods for Low-Light Scenarios Based on Dual-Stream Interleaved Fusion

Songhao Li
School of Computer Science and Engineering
Xi'an Technological University
Xi'an China
E-mail: glow365@163.com

ZhongSheng Wang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: wzhsh1681@163.com

***Abstract*—Object detection in low-light environments poses a challenging task. Existing work involving visible and infrared light primarily focuses on fusion enhancements across multiple scales of the backbone network. This paper proposes and constructs the Prior Difference Interleaved Network (PDIN), a dual-stream detector based on the YOLOv11n framework. Its core innovation lies in the interleaved deep fusion strategy combining CPCA and SDI. Specifically, the model introduces a Channel-Prior Convolutional Attention (CPCA) module before fusion to pre-enhance and reduce redundancy in dual-modal features. Subsequently, a Semantic Difference Interaction (SDI) module is designed and proposed, whose core lies in converting semantic differences between modalities into dynamic weight signals that guide fusion, achieving difference-driven adaptive integration. By first optimizing feature quality and then performing dynamic difference-driven integration, PDIN significantly enhances model robustness. Extensive results on the VEDAI dataset demonstrate PDIN's effectiveness, ultimately improving the mAP50 performance metric from the baseline 47.2% to 52.3%. This study robustly validates the efficacy of explicitly leveraging modal differences and performing feature quality pre-enhancement in bimodal deep learning fusion.**

Keywords-Object Detection, Feature Fusion, Visible-Infrared Light, Dynamic Weighting, Semantic Differences

I. INTRODUCTION

Thanks to advancements in sensing hardware, dual-modality target detection utilizing both visible light and infrared light has found widespread application in security surveillance, aerial remote sensing, and autonomous driving. Image in visible light reveal target details about texture and color, but are highly susceptible to loss of effective information under adverse conditions for example, in low light or heavy fog. Infrared images, based on thermal radiation imaging, can reliably output target contours in low-visibility scenarios but lack semantic information, making precise target classification difficult. Due to their significant differences, leveraging the supplementary information provided by infrared and visible light to support higher-level visual tasks remains challenging. support higher-level visual tasks remains challenging.

Multispectral image fusion methods like pixel-level weighting and feature channel concatenation often result in information loss and modal imbalance. Feature extraction and fusion across both spatial and channel dimensions yield superior detection results. Most existing visible and infrared detection models typically merge channel features after constructing a four-channel input, without pre-enhancing channel information or filtering redundant channels. This results in inefficient cross-modal information utilization. Additionally, the lack of dynamic capture capabilities for modal difference information can obscure crucial complementary details.

To resolve this issue, this paper proposes a Prior Difference Interleaved Network based on the lightweight YOLOv11 detection architecture. This network innovatively designs two core modules: First, the Channel A-priori Convolutional Attention module pre-enhances dual-modal features before fusion. It leverages channel a-priori

information to select high-discriminative channels and constructs a decoupled spatial attention mechanism through multi-scale deep convolutions. This generates independent dynamic attention maps for each channel, achieving feature redundancy reduction and activation of key regions. Second, the semantic difference interaction module generates adaptive interaction weights by encoding semantic distances between bimodal features. This drives network to focus on complementary areas, enabling difference-oriented dynamic fusion. Experimental results indicate on the vedai dataset, the PDIN model demonstrated outstanding detection performance. The proposed model builds upon the YOLOv11 framework, introducing innovative improvements in two dimensions: feature enhancement and dynamic fusion, specifically as follows:

A. We propose an SDI module leveraging modal difference mechanisms. By calculating and encoding the semantic distance between visible light and infrared features, it generates spatially adaptive interaction weight maps. This enables the network to automatically perceive and focus on complementary regions where differences are most pronounced.

B. Feature pre-enhancement is performed separately before channel fusion. Channel prior information first selects high-discriminative channels, while multi-scale deep convolutions build a decoupled spatial attention mechanism. This independently generates dynamic spatial attention maps for each feature channel, achieving feature redundancy reduction and focusing on critical regions.

II. RELATED WORK

This section first reviews single-modal object detection algorithms. Subsequently, it systematically summarizes recent work by object detection research teams on multimodal feature fusion and deep learning attention mechanisms.

A. Single-Modality Object Detection

In recent years, convolutional neural networks (CNNs) and their variants have been employed to improving the accuracy of object recognition tasks. Examples include SPP-NET, Sparse R-CNN, Fast R-CNN and YOLO series [1]. Subsequently, Carion et al. introduced the transformer-based detection transformer DETR [2]. Despite their impressive performance, these algorithms struggle with infrared images lacking texture. Researchers then turned to infrared image detection tasks, with Li et al. proposing YOLO-FIR [3] to enhance detection performance through channel compression, parameter optimization, and an improved attention module. IRSTD-GAN introduced a novel detection paradigm based on generative adversarial networks (GANs), focusing on the fundamental features of small infrared targets. It treats small infrared targets as specific noise and makes predictions based on data and features trained using GAN [4]. These models extract information solely from mono-modality images, yielding suboptimal results in complex scenes.

B. Multimodal Object Detection

Multimodal fusion based on visible and infrared light is a current research hotspot. Early on, popular methods for multimodal detection involved manually extracting features using HOG or ACF combined with SVM. Subsequently, Konig et al. were use RPN to fusion detection of infrared and visible images firstly. By integrating mid-level features from both modalities, their model could simultaneously reflect the characteristics of the two different imaging modes—RGB and IR—as learned through study. To overcome the limitations of manually designed fusion rules, researchers developed end-to-end image fusion algorithms based on convolutional neural networks (CNNs). Zhang et al. proposed a fast unified image fusion network method that preserves gradients and brightness ratios. enabling end to end fusion for multiple image modalities [5]. Xu et al. presented a unified unsupervised network for image fusion that automatically calculates t the weight of matched source images and acquire adaptive feature information The trained network maintains adaptive similarity between fused images and input source images [6]. However, CNN-based approaches are highly vulnerable to image misalignment due to intrinsic structural limitations. in handling interactions between

features within local regions. To address the deficiency in cross-modal long-range dependency modeling capacity of CNNs, researchers introduced Transformers into multispectral object detection. For instance, the CFT approach captures long-range dependencies during feature extraction and incorporates global contextual information under the guidance of a Transformer architecture [7]. Inspired by these works, this paper focuses more on the differences between the two modalities. By computing channel-wise difference information, dynamic weights are generated to leverage the complementary information among various modalities.

C. Attention Mechanism

Attention mechanisms have been broadly utilized within the realm of computer vision. Transformer-based ensemble methods primarily adopt or refine their self-attention mechanisms. SA-UNet adopts a spatial attention module that infers attention maps across the spatial dimension and multiplies these maps with the input feature maps to realize adaptive feature refinement [8]. TransAttUnet proposes a global spatial attention module that leverages global information modeling while utilizing self-attention mechanisms [9]. These techniques focus solely on spatially meaningful regions, neglecting attention to key objects in the channel dimension. SENet introduces a simple yet universal architecture by explicitly modeling inter-channel dependencies to adaptively recalibrate feature responses at the channel level [10]. CBAM is another lightweight, universal module that adaptively optimizes features across both spatial and channel dimensions [11]. ECANet employs a local cross-channel interaction strategy, achieving significant performance gains with minimal parameters through adaptive 1D convolutions [12]. The recently proposed CANet decodes neural network outputs into category-relevant features using category-specific dictionary learning, enabling the encoding of category attention maps [13]. CPCA proposes an efficient channel-prior convolutional attention method supporting weights to be dynamically distributed across channel and spatial dimensions [14]. By adopting multi-scale deep convolutional modules, it efficiently captures spatial relations while keeping channel priors intact. Consequently, CPCA possesses the capability to concentrate on informative channels and critical areas.

III. METHOD

This paper proposes the Priori Difference Interleaved Network (PDIN), a dual-modal object detection model built upon the YOLOv11 framework. The model employs a dual-stream feature extraction architecture and embeds channel-wise prior convolutional attention modules at multiple scale levels within the backbone network to pre-enhance and recalibrate dual-modal features. Subsequently, a semantic difference interaction module is introduced to perform difference-driven dynamic fusion, overcoming the inherent flaw in existing YOLO architectures where information in dual-modal fusion is often overwhelmed by strong feature information and background noise, making it difficult to efficiently extract and utilize complementary information.

A. Overview of the Overall Network Architecture

This paper proposes a dual-stream deep fusion model based on the advanced YOLOv11 framework. The model embeds the Channel-Priority Convolutional Attention (CPCA) and Semantic Difference Interaction (SDI) modules into the backbone network, forming a unique CPCA → SDI interleaved deep fusion strategy. Two parallel branches independently extract features from both visible and infrared modalities, and perform alternating intra-modal feature enhancement and cross-modal differential fusion across multiple spatial scales. While retaining the native architecture's advantages of lightweight design and real-time performance, the model effectively addresses feature redundancy and low complementary information utilization in existing algorithms, significantly improving detection accuracy and robustness for weak features, occluded objects, and small targets in real-world complex scenes. The overall structure of the proposed network is given in Figure 1 below:

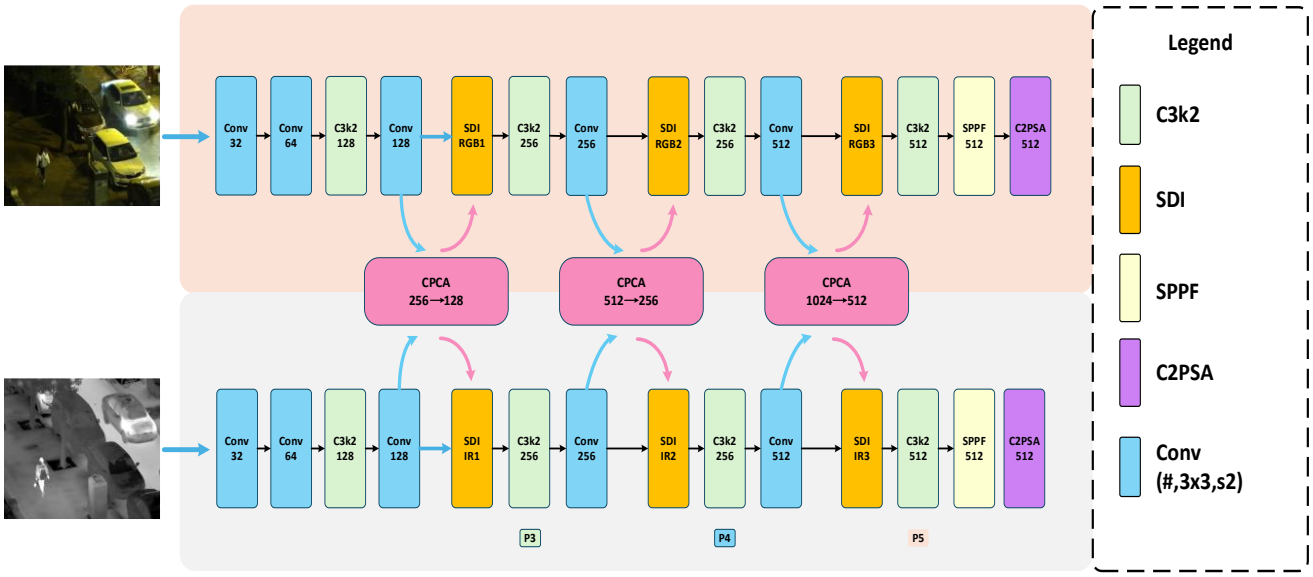


Figure 1. Overall Network Architecture Diagram

Feature extraction in the proposed model begins with a symmetric dual-stream backbone network specifically tailored for bimodal learning. Aligned visible and corresponding infrared inputs are combined on the channel dimension by concatenation to form a unified 6-channel tensor, then separated into two independent feature streams via the dedicated initial layer. These streams undergo parallel processing through multiple layers of Conv and C3k2 structures to progressively extract robust hierarchical multi-scale features for detection. At each scale, modal features first enter the CPCA module for pre-enhancement and redundancy reduction, leveraging its dynamic bidimensional attention to optimize feature quality and discriminative power while suppressing complex interfering background noise. Subsequently, the CPCA-enhanced features enter the SDI module, performing a difference-driven dynamic convex combination fusion to effectively maximize the utilization of valuable complementary information across modalities. The Neck structure employs classic PANet, efficiently aggregating rich high-level semantic and low-level detailed information across scales through dual top-down and bottom-up pathways. After multi-level bidirectional feature aggregation and propagation, the enhanced feature maps are fed into the YOLOv11 detection head, which concurrently outputs accurate object localization and fine-grained classification predictions across three predefined scales, completing the end-to-end bimodal object detection task robustly in real-world challenging scenes.

B. emantic Difference Interaction Module (SDI)

The design philosophy of this proposed module is as follows: Given the substantial disparity in image feature information between paired infrared and visible light, we leverage the absolute difference between the two to compensate for information loss during bimodal fusion. This difference dynamically generates an interaction weight map W, which ultimately enables adaptive feature integration through a dynamic learnable convex combination mechanism based on these weights. This approach maximizes complementary advantages while curbing redundant content, significantly augmenting discriminative capability of fused feature components and robustness against severe modal imbalance. Single-modality models exhibit insufficient perception of modal disparity information, struggling to effectively address modal imbalance issues caused owing to intricate environmental elements such as extreme light and occlusion. To overcome this limitation, this paper introduces the Semantic Disparity Interaction (SDI) module. The SDI module architecture is illustrated in Figure 2:

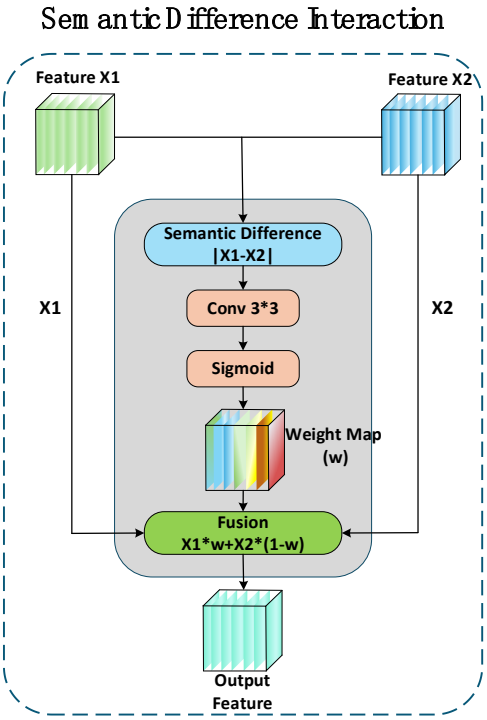


Figure 2. SDI Module Diagram

The SDI module receives visible light features $F_{rgb} \in R^{C \times H \times W}$ a and infrared features $F_{ir} \in R^{C \times H \times W}$ as input. The module first explicitly computes the absolute difference $| F_{rgb}-F_{ir} |$ between these two input features to highlight complementary and inconsistent information across the two modalities. The difference feature is then used as a guiding signal, processed through a lightweight a layer of 3×3 convolutional and a Sigmoid function σ, ultimately generating a dynamic interaction weight map W:

$$W = \sigma\left(Conv_{3x3}(| F_{rgb} - F_{ir} |)\right) \quad (1)$$

The generated weight map $W \in [0,1]$ possesses the same spatial dimensions and channel count as the input features. Each value in w represents the weight assigned to the visible light feature F_{rgb} at that position and channel. Finally, the SDI module employs the weight W to perform pixel-level and channel-level dynamic convex combination on F_{rgb} and F_{ir}, thereby ensuring the stability and effectiveness of the fused feature F_{fusion}:

$$F_{fusion} = W \odot F_{rgb} + (1-W) \odot F_{ir} \quad (2)$$

Here, ⊙ denotes element-wise multiplication. Since the range of W lies between [0,1], (1-W) automatically becomes the weight for the infrared feature F_{ir}. This adaptive weighting ensures that the fusion feature F_{fusion} dynamically favors the modality with richer or more stable information at both the pixel and channel dimensions, based on the level of difference across the two modalities features.

C. Channel Priority Attention Module (CPCA)

To further enhance feature discriminative power and optimize feature quality, this paper introduces the Channel Prior Convolutional Attention module. the CPCA module lies in supporting dynamic allocation of attention weights across both channel and spatial dimensions, enabling more precise focus on regions with maximum information content. Within this model, the CPCA module is embedded deep within the YOLOv11 backbone network—specifically at scales P3, P4, and P5—serving as a feature pre-enhancement module prior to SDI depth fusion. The CPCA module receives features from both visible and infrared branches. By combining channel-wise prior extraction with depth-wise convolution-based spatial modeling, it addresses the limited adaptability of traditional attention mechanisms. Through pre-alignment and redundancy reduction, it preemptively suppresses low-information channels and regions, passing more discriminative features to the subsequent SDI module. The architectural framework of the CPCA module is depicted in Figure 3:

The CPCA performs adaptive channel and spatial recalibration on the input bimodal features to output enhanced features Fout. Its implementation follows a dual-path attention mechanism, dynamically decoupling channel and spatial weights through deep convolutions.

The channel attention mechanism explicitly models the importance relationships among feature channels. First, The operations of global average pooling and global max pooling are performed in parallel on the input feature F to aggregate spatial information and generate two channel descriptors. These two descriptors are then introduced into a shared multi-layer

perceptron (MLP), processed through element-wise summation and sigmoid activation function σ, yielding the channel attention weights Mc:

$$M_c(F) = \sigma(MLP(AvgPool(F) + MLP(MaxPool(F))) \quad (3)$$

The core innovation of CPCA lies in generating the spatial attention Ms(F). To balance efficiency and adaptability, this paper avoids the spatial weight sharing constraints in CBAM by employing a multi-scale deep convolutional module to model spatial dependencies. Specifically, CPCA first concatenates pooled channel descriptors before feeding them into a Deep-Separable Convolution (DW-Conv) layer. This deep convolution independently processes each channel at low computational cost, enabling dynamic allocation of spatial weights across the channel dimension. Finally, after Sigmoid activation, the spatial attention weights Ms(F) are obtained:

$$M_s(F) = \sigma(DW - Conv(Concat(AvgPool(F), MaxPool(F)))) \quad (4)$$

The final augmented feature Fout is obtained by sequentially recalibrating the input feature F through both channel-wise and spatial attention mechanisms. The tandem recalibration approach ensures that features can more precisely focus on important spatial regions within channels that exhibit stronger discriminative power:

$$F' = M_c(F) \odot F \quad (5)$$

$$F_{out} = M_s(F) \odot F' \quad (6)$$

Among these, F' represents the intermediate features calibrated by channel attention. This tandem structure ensures that the final output Fout possesses robust feature discriminative power, providing high-quality, redundancy-free feature inputs for the subsequent SDI dynamic fusion module.

D. Evaluation Indicators

Recall Rate (abbreviated as R): Recall measures a model's ability to successfully detect all true targets. It represents the proportion of true targets. It indicates the ratio of true targets accurately recognized by the model. In object detection, a detection result is only considered a true positive (TP) is defined when IoU between target and ground truth exceeds a set threshold. The calculation formula is shown in Equation 7. A

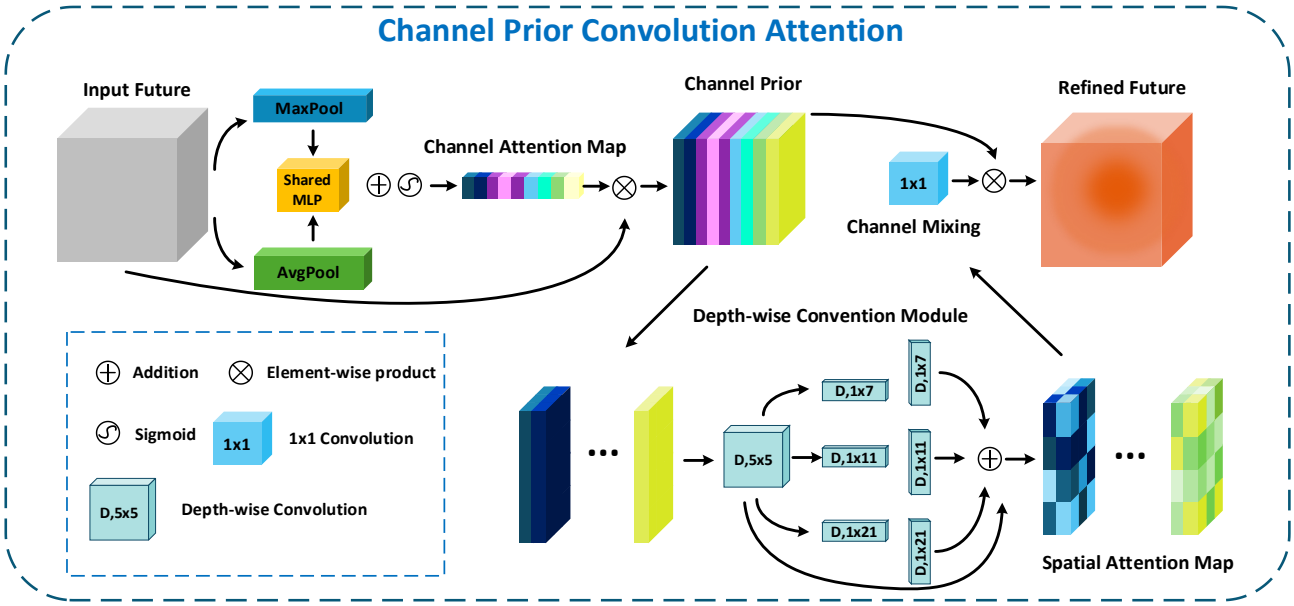


Figure 3. CPCA Block Diagram

common threshold value is 0.5.

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

TP denotes the count of correctly detected targets, while FN represents targets undetected by the model.

1) Accuracy rate (abbreviated as P): Accuracy measures the proportion of detected objects that are actual targets. It represents the percentage of all bounding boxes identified as targets by the model that are indeed real targets. The computational formula is given in Figure 8.

$$Precison = \frac{TP}{TP + FP} \tag{8}$$

The higher the P-value, the stronger the reliability of the model's recognition. The fewer misclassifications.

2) Average Precision (abbreviated as AP): Average Precision serves as an integrated metric to measure single-category detection capability. In object detection, as the confidence threshold decreases, recall increases but precision decreases. The calculation formula is as shown in 9.

$$AP = \frac{1}{n}\sum_{i=1}^{n} Precison_i \tag{9}$$

3) Mean Average Precision (abbreviated as mAP): The mean average precision acts as the ultimate evaluation index for overall performance of the model. It denotes the average of AP values for all categories in the dataset, and its computational formula is presented in Figure 10.

$$mAP = \frac{1}{|Q_R|}\sum_{q \in Q_R}^{n} AP(q) \tag{10}$$

4) Cross-Entropy Loss Function (abbreviated as Cross-Entropy Loss): In category prediction task of object detection, The cross-entropy loss is adopted as the core benchmark for assessing the discrepancy between the model-predicted class distribution and the real label distribution. It is commonly employed for training the category classification branch of detection boxes.

The model outputs a probability distribution $y = \left[y_1, y_2, ..., y_c \right]$ for C classes to each detection box. The actual category's unique heat encoding is $y = \left[y_1, y_2, ..., y_c \right]$. The single-sample cross-entropy loss calculation formula is as shown in 11:

$$L_{CE} = -\sum_{c=1}^{C} y_c \cdot log(y_c) \tag{11}$$

During batch training, the mean loss of N samples within a batch is taken as the overall category loss, as shown in Equation 12:

$$L_{CEbatch} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} \cdot \log(y_{i,c}) \tag{12}$$

Here, yi, c represents the ground-truth label for class c in the i-th sample, while $\hat{y}_{i,c}$ denotes the corresponding predicted probability. smaller loss value indicates higher accuracy in the model's class prediction. This loss is often combined with bounding box regression loss to form the overall loss for object detection.

E. PDIN Process Example

The forward propagation process of the proposed Prior Difference Interleaved Network (PDIN) is shown in Algorithm 1.

Algorithm 1 Overall Forward Propagation Process of PDIN Network

Input: Visible Image (Iv),Infrared Image (Ii)

Output: Detection Result

$\{V_{p3}, V_{p4}, V_{p5}\} \leftarrow Backbone\_Vis(I_v)$

$\{I_{p3}, I_{p4}, I_{p5}\} \leftarrow Backbone\_Inf(I_i)$

For $k \in \{p3, p4\}$:

    $Mattn \leftarrow CPCA(V_k, I_k)$

    $V_k \leftarrow C3k2(SDI(V_k, Mattn))$

    $I_k \leftarrow C3k2(SDI(I_k, Mattn))$

$Mattn\_5 \leftarrow CPCA(V_{p5}, I_{p5})$

$V_{p5} \leftarrow C2PSA(SPPF(SDI(V_{p5}, Mattn\_5)))$

$I_{p5} \leftarrow C2PSA(SPPF(SDI(I_{p5}, Mattn\_5)))$

$F_{out} \leftarrow \left[Concat(V_k, I_k) / k \in \{p3, p4, p5\} \right]$

Return $DetectionHead(F_{out})$

The algorithm first defines the input as visible light images and infrared images, with the output being detection results. (1-2) represents the dual-stream feature extraction phase, where features are extracted separately from two different images. (3-6) Cross-modal attention fusion is performed in the first and second stages (P3 and P4) using CPCA to generate attention maps. Subsequently, SDI and C3k2 are employed in update the visible and infrared feature maps, enhancing the fusion features across both modalities. (7-9) Deep Fusion Stage: This section describes the third stage (P5), where cross-modal attention fusion is performed via CPCA to generate an attention map. Subsequently, SDI processes the visible and infrared feature maps, followed by further enhancement of the fused features through SPPF and C2PSA. (10-11) Describes the concatenation of visible and infrared feature maps from the above three stages to obtain fused features, followed by final object detection predictions through the detector head.

IV. Experiments and Analysis

This section aims to validate the effectiveness of the proposed Semantic Difference Interaction (SDI) module and Channel-Prior Convolutional Attention (CPCA) module through systematic experiments. We first introduce the detailed dataset and experimental setup, followed by ablation studies to quantitatively analyze the contributions and synergistic effects of each module. Finally, we compare the final model obtained from our research with existing methods to establish its advantages within the scope of bimodal object detection.

A. Dataset

This study employs VEDAI dataset to validate effectiveness of the proposed model. The VEDAI dataset serves as a common benchmark within the scope of aerial target detection, comprising visible light images and their paired infrared images, thereby providing a natural testing environment for bimodal fusion. The dataset contains targets across nine categories, including automobiles, trucks, and vessels, with image scenes encompassing complex ground environments and lighting conditions. This study selected valid data pairs with aligned labels from the dataset and split into training, validation and test sets with a 3.5:1:1 ratio. Dataset statistics are summarized in Table 1 below, while Figure 4 illustrates the sample distribution across categories within the dataset.

TABLE I. VEDAI DATASET

Category	Image pair	Number of annotations
Train_RGB	874	874
Trsin_IR	874	874
Val_RGB	248	248
Val_IR	248	248
Test	248	248

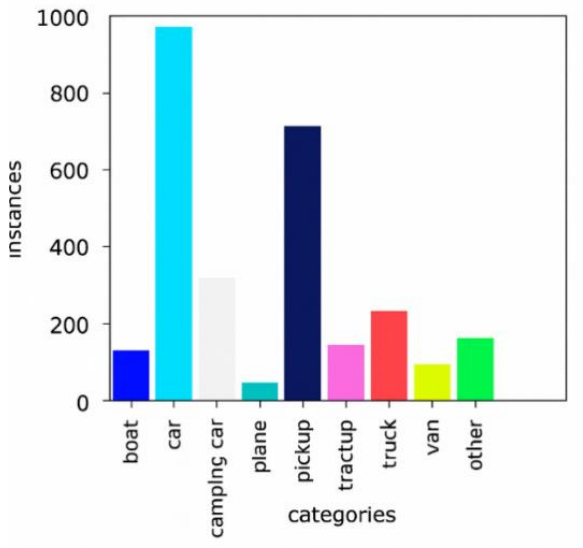


Figure 4. Dataset Categories

B. Training Settings

In this experiment, We adopted the AdamW optimizer for model training. with an initial learning rate η set to 0.002. The model was trained for a total of 200 epochs. We uniformly configured the batch size to 4, and all experiments were performed using the PyTorch framework on a machine installed with a single NVIDIA 3080 Ti GPU, with workers set to 8. Additionally, automatic mixed precision (AMP) was disabled (AMP=False) across all experiments to avoid potential compatibility issues. Detailed environment information is provided in Table 2:

TABLE II. ENVIRONMENT CONFIGURATION

Name	Model
Processor	12vCPUIntel(R)Xeon(R)Silver4214RCPU@2.40GHz
Graphics card	RTX3080Ti(12GB)*1
Memory	90GB
Hard disk	50GB
System	Linux

C. Training Results

The training and validation metric curves of the model are shown in Figure 5.

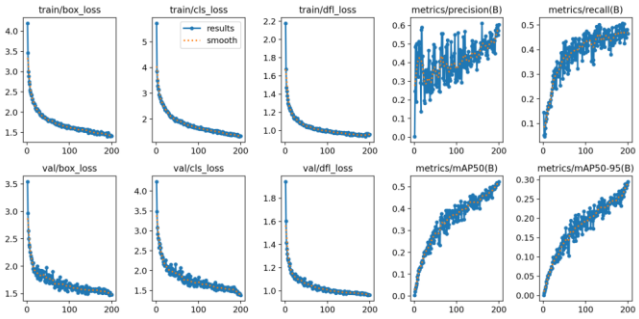


Figure 5. Training and Validation Metric Curves

The entire training process totaled 200 epochs. All loss curves—including train/box_loss, train/cls_loss, and train/dfl_loss all exhibit stable and rapid convergence trends without significant oscillations or divergence. This indicates that the optimizer and hyperparameter settings employed for model training are stable and effective. On the validation set, all losses rapidly decreased initially before entering a plateau phase and maintaining low values, indicating that the model did not exhibit severe overfitting on the validation set. Concurrently, the key performance metrics/mAP50(B) and metrics/mAP50-95(B) steadily increased with training iterations, with the PDIN model achieving the highest performance.

Figure 6 displays the precision-recall (PR) curve of the PDIN model on the VEDAI test dataset. The thick blue line signifies the average PR curve across all categories, with area under this curve serving as the core metrics/mAP@0.5, ultimately settling at 0.523. This aligns with the optimal result obtained in the ablation experiments (Table 4) presented in this paper. The overall trend of this curve indicates that the model achieves high recall at the cost of a slight reduction in precision. Figure 7 below illustrates the detection results of the PDIN model.

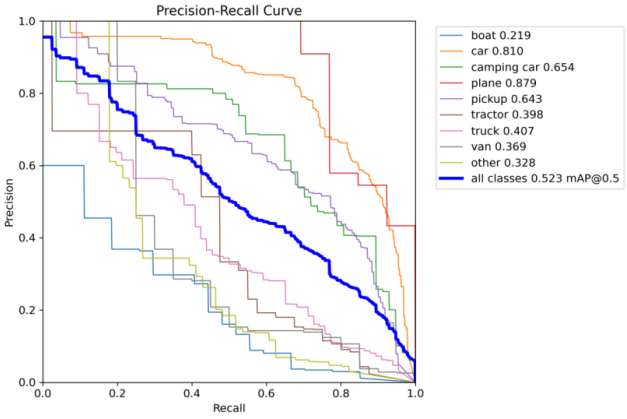
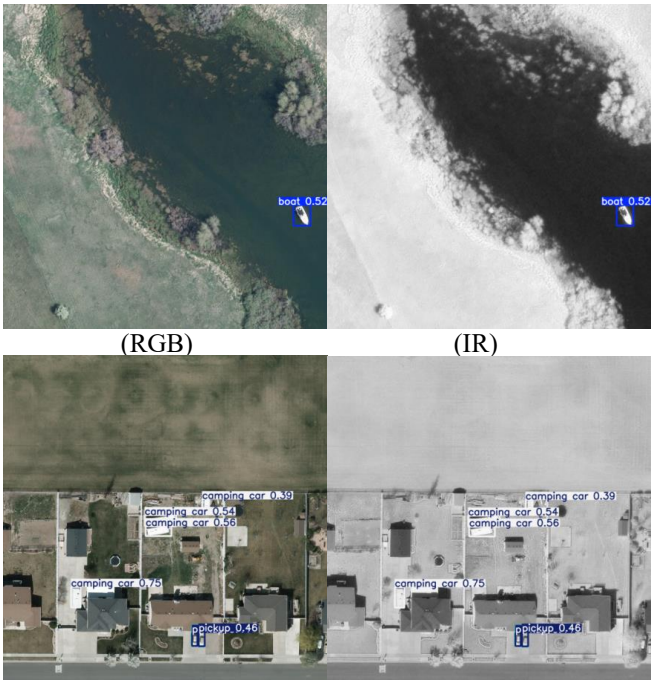


Figure 6. Precision-Recall (PR) Curve



(RGB) (IR)

(RGB) (IR)

Figure 7. Test results

D. Melting Experiment

This section aims to assess the effectiveness and optimal configuration of the proposed Semantic Difference Interaction (SDI) module and Channel-Prior Convolutional Attention (CPCA) module by conducting a set of experiments. All experiments use a dual-stream model based on the YOLOv11n architecture as the baseline, with mAP50 as the core evaluation metric.

This subsection investigates the effect of SDI module count on performance. Experiments were conducted by sequentially increasing the number of SDI modules within the backbone network while keeping the CPCA count constant. Experimental data are presented in Table 3 below. The results indicate that model performance peaks when the number of SDI modules is increased to 3.

TABLE III. THE IMPACT OF MODULE QUANTITY ON MODEL PERFORMANCE

Number of blocks	mAP-50	mAP50-95
1	0.48521	0.28444
2	0.51255	0.27454
3	0.52345	0.29499
4	0.50912	0.29186

red: The best indicator.

We aim to quantitatively evaluate the effectiveness of the proposed Semantic Disparity Interaction (SDI) module and Channel-Prior Convolutional Attention (CPCA) module, a series of ablation experiments were designed. All experiments employed a dual-stream model based on the YOLOv11n architecture as the baseline, progressively adding the SDI and CPCA modules for performance testing. We adopted mAP50 as the major evaluation metric, with relevant experimental data illustrated in Table 4.

TABLE IV. ABLATION EXPERIMENT RESULTS FOR SDI AND CPCA MODULES

Index	Model	precision	recall	mAP-50	mAP50-95
1	Baseline	0.52726	0.49901	0.47278	0.26334
2	+SDI only	0.55986	0.55833	0.48523	0.27637
3	+CPCA only	0.56296	0.54735	0.51259	0.28799
4	+SDI&CPCA	0.60456	0.55752	0.52345	0.29499

red: The best indicator.

Analysis of the results shows that training based on Dual-Stream YOLOv11n achieved a mAP50 score of 47.2%. When the SDI module was designed and introduced into the base model alone, the model performance saw a significant increase of 1.3 percentage points. This outcome demonstrates the effectiveness of the Semantic Differentiated Interaction strategy in handling multi-modal information. Introducing only the CPCA module into the base model improved performance by 4.0 percentage points, highlighting how CPCA's dynamic dual-dimensional attention mechanism greatly improves the model's capability to extract crucial feature information after feature redundancy reduction and recalibration. Ultimately, integrating both SDI and CPCA modules into the model yielded the highest performance, achieving a mAP50 score of 52.3%. The PDIN model demonstrated a total promotion of 5.1 percentage points relative to the baseline model. This outcome fully demonstrates that feature quality enhancement and dynamic complementary information integration reinforce the model's resilience in complex environments.

E. Comparative experiment

This study employs a rigorous variable control method and utilizes the current mainstream cross-modality fusion technology (Cross-Modality Fusion Transformer), hereinafter referred to as CFT. Similarly based on the YOLOv11 model, it modifies the dual-stream architecture for infrared and visible light, maintaining consistent parameters with previous work. The Vedai dataset is again selected, with training conducted for 200 epochs. Specific results are presented in the table 5:

TABLE V. PERFORMANCE COMPARISON WITH CFT FUSION TECHNOLOGY

Method	precision	recall	mAP-50	mAP50-95
Yolov11+CFT	0.57867	0.54634	0.42978	0.23415
Yolov11+SDI	0.55986	0.55833	0.48523	0.27637
Yolov11+CFT+CPCA	0.56158	0.45254	0.42740	0.23568
Yolov11+SDI+CPCA	0.60456	0.55752	0.52345	0.29499

red: The best indicator.

The proposed PDIN interleaved deep fusion network demonstrates outstanding performance in the vedai bimodal object detection task. Quantitative comparisons reveal that the difference-driven SDI module (mAP50=0.48523) significantly outperforms the globally dependency-based CFT module (mAP50=0.42978), validating the effectiveness of the SDI strategy in addressing modal imbalance. The final integrated model combining CPCA pre-enhancement and SDI dynamic fusion achieves the

highest detection rate (mAP50=0.52345), representing a substantial 9.37 percentage point improvement over mainstream CFT methods. Qualitative analysis further demonstrates that our approach effectively handles complex scenes, small objects, and low-visibility challenges, exhibiting higher detection accuracy and robustness in engineering applications.

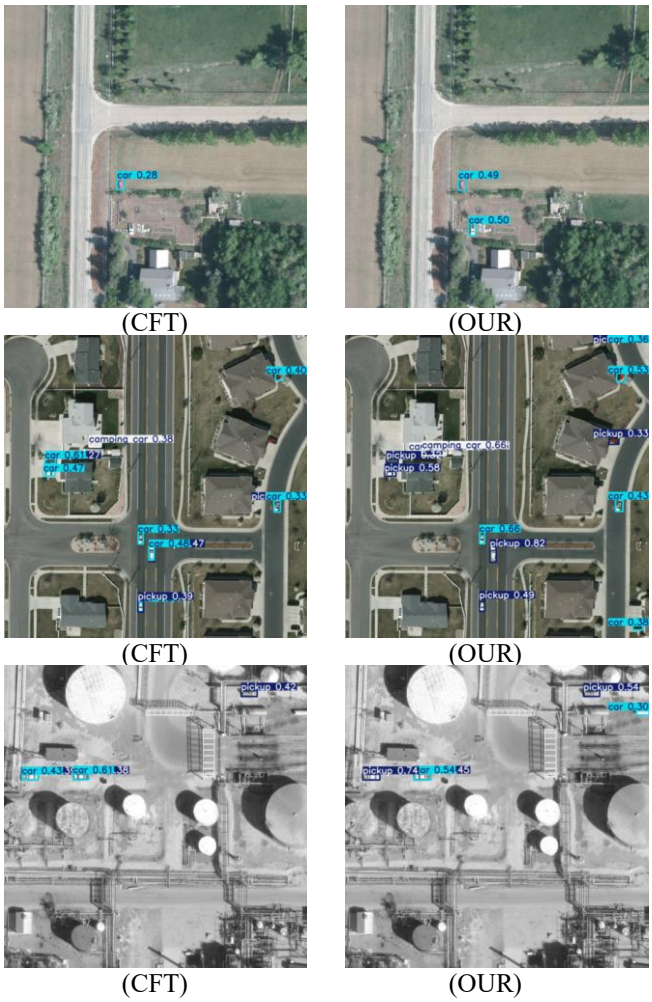


Figure 8. Comparison of Detection Results Between CFT and PDIN Models

To more straightforwardly reflect this model's superiority in practical engineering applications, we performed object detection tasks on untrained test sets using the optimal weights obtained from training. Figure 8 presents comparative detection results between our method and CFT under challenging conditions including low visibility, small targets, and complex scenes. These representative images clearly reveal significant improvements in our model's false alarm rate, miss detection rate, and localization accuracy.

V. CONCLUSIONS

The proposed Prior Difference Interleaved Network (PDIN) achieves stepwise, refined modal information integration across multi-scale levels of the YOLOv11 backbone through an interleaved deep fusion strategy combining Channel Prior Convolutional Attention (CPCA) and Semantic Difference Interaction (SDI). Ultimately, the PDIN model demonstrates significant performance improvements on the VEDAI dataset. Ablation experiments demonstrate that the model achieves a 5.1 percentage point improvement over baseline models on the core evaluation metric mAP50. Furthermore, PDIN exhibits superior performance compared to mainstream CFT approaches. This research holds broad application prospects in engineering domains. Addressing computational constraints on UAVs and edge computing devices, subsequent work will be dedicated to lightweight deployment of the PDIN to reduce computational overhead associated with bimodal fusion.

REFERENCES

[1] Sun P, Zhang R, Jiang Y, et al. Sparse r-cnn: End-to-end object detection with learnable proposals[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 14454-14463.

[2] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 213-229.

[3] Li S, Li Y, Li Y, et al. Yolo-firi: Improved yolov5 for infrared image object detection[J]. IEEE access, 2021, 9: 141861-141875.

[4] Zhao B, Wang C, Fu Q, et al. A novel pattern for infrared small target detection with generative adversarial network[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 59(5): 4481-4492.

[5] Zhang H, Xu H, Xiao Y, et al. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(07): 12797-12804.

[6] Xu H, Ma J, Jiang J, et al. U2Fusion: A unified unsupervised image fusion network[J]. IEEE transactions on pattern analysis and machine intelligence, 2020, 44(1): 502-518.

[7] Qingyun F, Dapeng H, Zhaokui W. Cross-modality fusion transformer for multispectral object detection[J]. arXiv preprint arXiv:2111.00273, 2021.

[8] Guo C, Szemenyei M, Yi Y, et al. Sa-unet: Spatial attention u-net for retinal vesselsegmentation[C]//2020 25th international conference on pattern recognition (ICPR). IEEE, 2021: 1236-1242.

[9] Wang W, Chen C, Ding M, et al. Transbts: Multimodal brain tumor segmentation using transformer[C]//International conference on medical image computing and computer-assisted intervention. Cham: Springer International Publishing, 2021: 109-119.

[10] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[11] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[12] Wang Q, Wu B, Zhu P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11534-11542.

[13] Cheng G, Lai P, Gao D, et al. Class attention network for image recognition[J]. Science China Information Sciences, 2023, 66(3): 132105.

[14] Khanam R, Hussain M. Yolov11: An overview of the key architectural enhancements[J]. arXiv preprint arXiv:2410.17725,2024.

[15] Fu H, Wang S, Duan P, et al. Lraf-net: Long-range attention fusion network for visible–infrared object detection[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 35(10): 13232-13245.

[16] Sharma M, Dhanaraj M, Karnam S, et al. YOLOrs: Object detection in multimodal remote sensing imagery[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 14: 1497-1508.

[17] Shen J, Chen Y, Liu Y, et al. ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection[J]. Pattern Recognition, 2024, 145: 109913.

[18] Chen Y, Wang B, Guo X, et al. DEYOLO: Dual-feature-enhancement YOLO for cross-modality object detection[C]//International Conference on Pattern Recognition. Cham: Springer Nature Switzerland, 2024: 236-252.

[19] Zhong J, Zhang J. MIR-YOLO: Remote sensing small target detection network based on visible-infrared dual modality[J]. Digital Signal Processing, 2025, 162: 105158.

[20] Li X, Yan H, Cui K, et al. A novel hybrid YOLO approach for precise paper defect detection with a dual-layer template and an attention mechanism[J]. IEEE Sensors Journal, 2024, 24(7): 11651-11669.