FD^2-Net: Frequency-Driven Feature Decomposition Network for Infrared-Visible Object Detection

Ke Li[1], Di Wang[1*], Zhangyuan Hu[1], Shaofeng Li[1*], Weiping Ni[2], Lin Zhao[3], Quan Wang[1]

[1]Xidian University,
[2]Northwest Institute of Nuclear Technology,
[3]Nanjing University of Science and Technology

Abstract

Infrared-visible object detection (IVOD) seeks to harness the complementary information in infrared and visible images, thereby enhancing the performance of detectors in complex environments. However, existing methods often neglect the frequency characteristics of complementary information, such as the abundant high-frequency details in visible images and the valuable low-frequency thermal information in infrared images, thus constraining detection performance. To solve this problem, we introduce a novel **F**requency-**D**riven **F**eature **D**ecomposition **Net**work for IVOD, called FD^2-Net, which effectively captures the unique frequency representations of complementary information across multimodal visual spaces. Specifically, we propose a feature decomposition encoder, wherein the high-frequency unit (HFU) utilizes discrete cosine transform to capture representative high-frequency features, while the low-frequency unit (LFU) employs dynamic receptive fields to model the multi-scale context of diverse objects. Next, we adopt a parameter-free complementary strengths strategy to enhance multimodal features through seamless inter-frequency recoupling. Furthermore, we innovatively design a multimodal reconstruction mechanism that recovers image details lost during feature extraction, further leveraging the complementary information from infrared and visible images to enhance overall representational capacity. Extensive experiments demonstrate that FD^2-Net outperforms state-of-the-art (SOTA) models across various IVOD benchmarks, *i.e.* LLVIP (96.2% mAP), FLIR (82.9% mAP), and M^{3}FD (83.5% mAP).

Introduction

Object detection is a foundational topic in computer vision, aiming to localize and identify diverse objects within images or videos. It has extensive applications in autonomous driving, surveillance, and remote sensing (Fu et al. 2023b; Li et al. 2023b). Nevertheless, visible object detection encounters substantial challenges in adverse conditions like rain, fog, clouds, and poor illumination, primarily due to the inherent limitations of RGB sensors. As a result, alternative visual sensors, particularly infrared cameras, are increasingly utilized to complement RGB cameras in overcoming these difficulties, thereby igniting substantial research interest in Infrared-Visible Object Detection (IVOD).

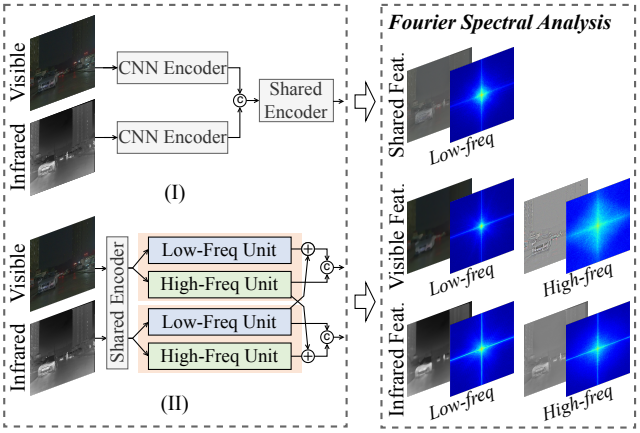

Figure 1: Illustration of the differences between our FD^2-Net and existing IVOD approaches. Our algorithm employs frequency decoupling to separate high- and low-frequency information in infrared and visible images, thereby effectively leveraging multimodal complementary features to extract more discriminative and robust characteristics.

However, current IVOD methods still have three weaknesses. ***Weakness 1:*** They tend to overlook the frequency characteristics of object features within infrared and visible images. Infrared imaging primarily captures low-frequency thermal radiation, while visible imaging emphasizes high-frequency details. Prevailing architectures (Li et al. 2023a; Zhao et al. 2023c) often overlook this kind of intrinsic property, and embed cross-modality information into a unified feature space, which results in the inability to extract modality-specific features. ***Weakness 2:*** With a fixed receptive field, these methods only extract local information, which makes it difficult to adapt to the positional biases inherent in infrared and visible images. Moreover, models with small kernels are inadequate for effectively capturing long-range information, which is crucial as the surrounding environment provides vital clues about object size, shape, and other characteristics (Li et al. 2024). ***Weakness 3:*** Recent IVOD approaches commonly employ downsampling operations to mitigate visual noise and reduce computational

*Corresponding author.

overhead, potentially resulting in the loss of object information. Such degradation in feature representation significantly hampers the localization and classification capabilities of the detection head, ultimately compromising detection performance. Our research explores a more rational paradigm to address these challenges in cross-modality feature extraction for IVOD tasks. Based on the aforementioned analysis, we identify three critical countermeasures (CM):

CM 1: We revisit the feature extraction process from a frequency perspective. Visible images furnish abundant high-frequency information, such as edges and textures, whereas infrared images deliver valuable low-frequency thermal radiation information. As illustrated in Fig. 1 (I), conventional methods depends solely on redundant cross-modality similar clues, leading to the loss of crucial complementary features. In contrast, we can capture discriminative complementary information from infrared and visible images in a more controlled and interpretable manner by limiting the frequency space of feature extraction. As shown in Fig. 1 (II), adaptive frequency decoupling facilitates the retention of more representative low-frequency and high-frequency information in both infrared and visible images.

CM 2: From a model design standpoint, larger kernels aid in capturing more extensive scene context, thereby mitigating geometric biases between infrared and visible images. However, employing large kernel convolutions may introduce substantial background noise and overlook fine-grained details within the receptive field, which can be detrimental to the precise detection of small objects. Hence, we parallelly arrange multiple depthwise dilated convolutions of varying sizes to extract multi-granularity texture features across diverse receptive fields, thus fulfilling IVOD tasks.

CM 3: To combat the information loss resulting from repeated downsampling, many existing methods often employ generative approaches such as image super-resolution to alleviate this issue. However, these methods not only require constructing pairs of high-resolution and low-resolution samples, but their generative processes often introduce spurious artifacts. Conversely, we integrate a simple yet effective multimodal reconstruction mechanism into the IVOD framework, leveraging complementary information from both infrared and visible modalities to restore structural and texture details lost during feature extraction.

In this paper, we design a novel paradigm for IVOD tasks, *i.e.*, *Frequency-Driven Feature Decomposition Network (**FD^2-Net**)*, which decouples the frequency information of infrared and visible images to efficiently extract representative features and leverages the dominant frequency characteristics of one modality to enhance the complementary features of the other. Specifically, we introduce a feature decomposition encoder, which comprises three main parts: a high-frequency unit (HFU), a low-frequency unit (LFU) and a parameter-free complementary strengths strategy (CSS). HFU performs the discrete cosine transform, followed by a lightweight module that learns a spatial attention mask from multiple high-frequency components, thereby accentuating the most representative high-frequency features. LFU employs multi-scale convolutional kernels to capture low-frequency structures of various objects and their contextual information, effectively modeling the relationships between objects and their surrounding environments. Subsequently, CSS leverages the strengths of one modality to achieve complementary enhancement in the other. Furthermore, we develop a cross-reconstruction unit (CRU) incorporating feature-level complementary masks. CRU further learns complementary information from infrared and visible features through both fine-grained and coarse-grained cross-modality interactions, restoring the multimodal images. Our contributions can be summarized as follows:

- We propose a novel paradigm for IVOD, termed FD^2-Net, which aims to improve detection performance by effectively extracting valuable complementary features from infrared and visible images.
- We design a high-frequency unit (HFU) and a low-frequency unit (LFU) to effectively capture discriminative frequency information in both infrared and visible images. Also, a complementary strengths strategy is developed to enhance multimodal features through seamless inter-frequency recoupling.
- We introduce a cross-reconstruction unit (CRU) to facilitate the fusion of complementary information across modalities, thereby further enhancing feature representation.
- Extensive qualitative and quantitative experiments validate the effectiveness of our FD^2-Net, achieving accuracies of 96.2% on LLVIP (Jia et al. 2021), 82.9% on FLIR (F.A. Group 2018), and 83.5% on M^{3}FD (Liu et al. 2022a).

Related work

General Object Detection

General object detectors can be broadly classified into two-stage detectors and one-stage detectors. Faster R-CNN (Ren et al. 2015) is a classic two-stage detector, consisting of a Region Proposal Network (RPN), Region of Interest (RoI) pooling, and detection heads. The RPN generates proposals based on features extracted by the backbone network. The extracted image features and generated proposals are fed into the RoI pooling operation to extract proposal features. Finally, the proposal features are classified and regressed by the detection head. To generate better region proposals, various methods have been explored to enhance performance, including architecture design (Cai and Vasconcelos 2018), anchor box optimization (Jiang et al. 2018), and multi-scale training (Singh, Najibi, and Davis 2018). However, two-stage methods necessitate filtering a large number of proposals, leading to significant time and computational overhead. In contrast, one-stage detection frameworks predict bounding boxes and classes directly from densely sampled grids, thus achieving faster inference speeds. YOLOv1 (Redmon et al. 2016) is the first one-stage object detector to achieve real-time object detection. Through years of continuous development, the YOLO detectors have surpassed other one-stage object detectors (Liu et al. 2016; Lin et al. 2017) and become synonymous with real-time object detection. In this article, YOLO-based architicture is chosen as the detector to reasonably balance speed and accuracy.

Infrared-Visible Object Detection

Infrared-visible fusion can complementarily capture richer object information, yielding more stable detection results. The main focus of IVOD detectors has primarily been on exploring improved fusion techniques, for which several variant frameworks have been proposed. TINet (Zhang et al. 2023d) enhances the extraction of complementary information by emphasizing the differences between infrared and visible images. AR-CNN (Zhang et al. 2019) highlights that visible images and infrared images are misaligned in the spatial dimension. To align the regional features of two modalities, it proposes a region feature alignment module to enhance detection performance. Furthermore, DMAF (Zhou, Chen, and Cao 2020) designs an illumination-aware feature alignment module that selects features based on illumination conditions and adaptively aligns features across modalities. To effectively capture the complementary features of infrared-visible images, APWNet (Zhang et al. 2023c) introduces an image fusion loss to enhance the performance of YOLOv5 (Jocher 2020). SuperYOLO (Zhang et al. 2023a) adds an image super-resolution branch to strengthen the feature extraction capability of the backbone. LRAF-Net (Fu et al. 2023b) improves detection performance by fusing the long-range dependencies of the visible and infrared features. DFANet (Zhang et al. 2023b) introduces an antagonistic feature extraction and divergence module to extract the differential infrared and visible features with unique information.

In this paper, we propose a frequency-driven feature decomposition network that can efficiently extract discriminative complementary information from infrared and visible images, respectively. This extracted information is then utilized to enhance feature representation, thereby improving detection performance.

Proposed Method

Overall Architecture

As shown in Fig. 2, our FD^2-Net comprises three modules: **1) Feature Decomposition Encoder.** Inspired by spectral spectrum, this module introduces a two-branch architecture to effectively extract valuable high-frequency and low-frequency features through feature decomposition and fusion. Subsequently, through the complementary advantage strategy, the representative frequency features are reorganized to improve the overall representation ability. **2) Multimodal Reconstruction Mechanism.** To enhance feature learning, an asymmetric cross-maske strategy is applied to the features from the final layer of the Encoder, compelling each modality to obtain useful information from complementary modalities. Two cross-reconstruction units are then used to restore the multimodal image by leveraging the complementary features of infrared and visible images. The reconstruction process is constrained by the mean square error at the pixel level. **3) Multi-Scale Detection Heads.** This module constructs a Feature Pyramid Network (FPN) that utilizes multi-scale features extracted at various stages of the Encoder. At the highest resolution layer of the FPN, the reconstructed multimodal features are integrated to further enhance detection. Finally, following YOLOv5 (Jocher 2020), three detection heads with different scales are configured to accurately detect objects.

Feature Decomposition Encoder

Formally, let $I \in \mathbb{R}^{H\times W}$ and $V \in \mathbb{R}^{3\times H\times W}$ be the input infrared and visible images, where $H \times W$ represents the spatial resolution. Initially, a 6×6 CBR block[1] is employed to reduce the resolution and extract shallow multimodal visual features $\{X_I^S, X_V^S\} \in \mathbb{R}^{c\times h\times w}$. Then, we first split $\{X_I^S, X_V^S\}$ into two components in a ratio of α, respectively. One is expected to represent high-frequency component, denoted as $\Phi^H = \{X_I^H, X_V^H\} \in \mathbb{R}^{\alpha c\times h\times w}$, capturing the spatial details such as edges and textures. The other $\Phi^L = \{X_I^L, X_V^L\} \in \mathbb{R}^{(1-\alpha)c\times h\times w}$ is expected to learn low-frequency content like context and structural information.

High-Frequency Feature Attention. Our goal is to effectively extract high-frequency components from infrared and visible images, respectively. We thus introduce a High-Frequency Unit, which can filter out the high-frequency information and direct model's attention on more valuable information. The discrete cosine transform (DCT) has demonstrated superiority in image compression, particularly in enhancing image details and textures while eliminating noise. Based on this, we incorporate DCT into IVOD. This transformation guides the convolution to extract diverse high-frequency spatial features and effectively suppresses noises such as Gaussian and thermal noise in infrared-visible images.

Discrete Cosine Transform (DCT). For an imgae $x \in \mathbb{R}^{H\times W}$, where H and W are the height and width of x, Eq. (1) provides the definition of the standard two-dimensional (2D) DCT, mathematically defined as:

$$f^{h,w} = \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} x_{i,j} B_{i,j}^{h,w}, \tag{1}$$

$$B_{i,j}^{h,w} = \cos(\pi h/H(i+1/2))\cos(\pi w/W(j+1/2)), \tag{2}$$

where $f \in \mathbb{R}^{H\times W}$ is the 2D DCT frequency spectrum, B is the basis function of the 2D DCT, $h \in \{0, H-1\}$ and $w \in \{0, W-1\}$, and $\cos(\cdot)$ represents the cosine function. To simplify the notation, constant normalization factors in Eq. (1) are omitted.

High-Frequency Unit. To adaptively modulate the emphasis on different frequency components for enhanced spatial information discrimination, we leverage the 2D DCT as a selective filtering mechanism. Specifically, the high-frequency feature maps Φ^H are divided along the channel dimension into n segments. Each group Φ_g^H, where $g \in \{0, n-1\}$, maintains the spatial dimensions of Φ^H but has only $1/n$ of channel length. A specific 2D DCT frequency component, denoted as B_{u_g,v_g}, is then assigned to each segment and then concatenate to obatain the modality-specific high-frequency features, which is denoted as:

$$\Phi^H = [\Phi_0^H * B_{u_0,v_0}, \cdots, \Phi_{n-1}^H * B_{u_{n-1},v_{n-1}}], \tag{3}$$

[1]A 6×6 convolutional layer with a batch normalization (BN) (Ioffe and Szegedy 2015) layer and a rectified linear unit (ReLU) (Nair and Hinton 2010).

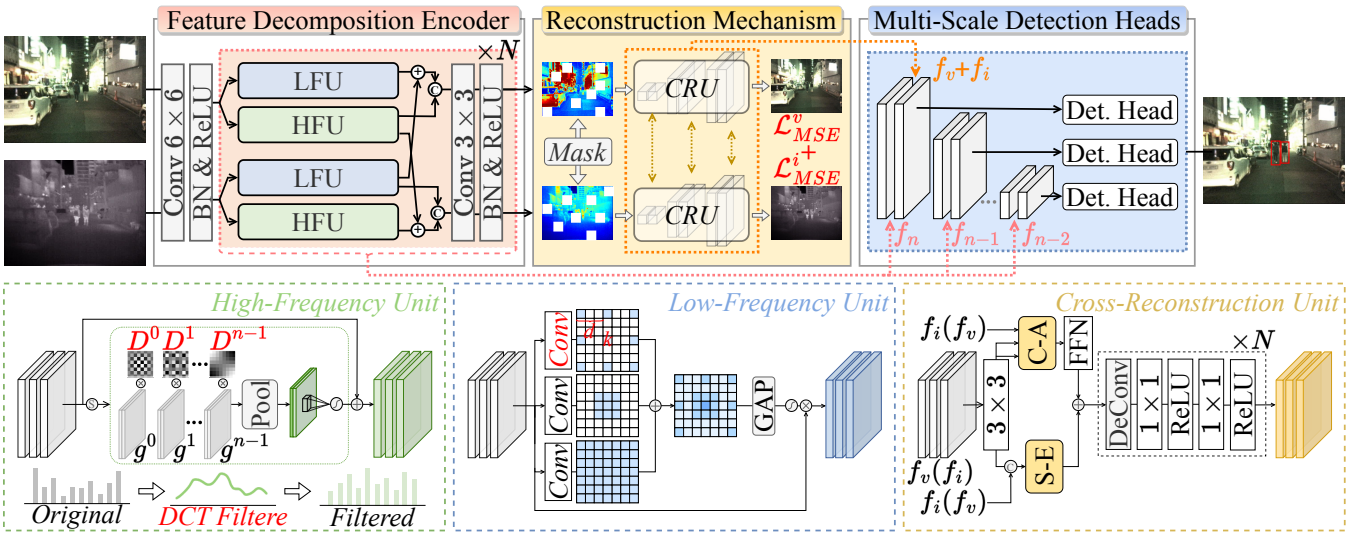

Figure 2: The architecture (top row) and core components (bottom row) of our FD^2-Net. It has three components: (1) Feature Decomposition Encoder, which effectively extracts high/low-frequency features in multimodal visual space. (2) Multimodal Reconstruction Mechanism, which further learns the distinguishing and complementary features of each modality through the reconstruction of multimodal images to enhance feature representation. (3) Multi-Scale Detection Head, which uses visual features from (1) and (2) to complete object classification and localization.

where $[u_g, v_g]$ represents the 2D frequency indices corresponding to Φ_g^H. The $[\cdot,\cdot]$ is a concatenation operation. Here, g serves as the control parameter for frequency components, where a larger g value enables channels within the same convolutional layer to capture multi-frequency features, thereby enhancing feature representation capability.

Next, we apply a spatial attention mechanism to adaptively learns a spatial mask to dynamically modulate different frequency components during training. Mathematically, this instantiation can be formulated as:

$$\mathrm{SA}^H = \sigma(\mathcal{F}_{7\times7}^{2\to1}[\mathrm{AvgPool}(\Phi^H), \mathrm{MaxPool}(\Phi^H)]), \quad (4)$$

where σ denotes the sigmoid function. AvgPool$(\cdot)$ and MaxPool$(\cdot)$ are average-pooling and max-pooling operations. $\mathcal{F}^{2\to1}$ is a 7×7 convolutional layer to transform the features (with 2 channels) into one spatial attention map, which facilitates information interaction among various spatial descriptors.

The final output of the HFU module is the element-wise product of the input feature Φ^H and SA^H, as indicated below:

$$\Phi^{H'} = \mathrm{SA}^{hf} \otimes \Phi^H. \quad (5)$$

We believe using more complex attention architectures, such as (Behera et al. 2021; Bao et al. 2024), is of the potentials to achieve higher improvements.

Low-Frequency Context Refinement. To effectively capture low-frequency information across multiple scales, we construct a multi-granularity convolution by a set of parallel depth-wise convolutions (DWCs) with different kernel sizes and dilation rates. For the i-th DWC, the expansion of the kernel size k_i and dilation rate d_i are flexible, with the only constraint being:

$$k_i + (k_i - 1) \times (d_i - 1) \leq RF. \quad (6)$$

However such a multi-branch structure invariably increases computational cost, thereby prolonging inference times in practical deployments. References (Ding et al. 2019, 2021) point out that multiple parallel convolutional blocks can be seamlessly consolidated into a single convolutional layer for inference, optimizing computational efficiency. By leveraging this equivalent transformation, we merge several small-kernel branches into a unified large-kernel convolutional layer, as shown in Fig. 2. This approach not only enhances the extraction of multi-scale features within a single layer but also maintains swift inference capabilities. Following the ConvNeXt (Liu et al. 2022b) and RepLKNet (Ding et al. 2022), we set $RF = 7$, kernels size is $[7, 3, 3, 3]$ and dilation rates is $[1, 1, 2, 3]$. Note that our LFU uses dilated convolution, thereby preventing the extraction of overly dense feature representations.

To diminish information redundancy and improve the feature diversity, we employ a channel-mix strategy that performs both inter-channel communications and spatial aggregations. First, a Global Average Pooling (GAP) operation collates channel statistics from low-frequency spatial features. These features then undergo compression and restoration via two sequential 1×1 convolution layers, reducing feature similarity. A sigmoid function subsequently generates channel weights, refining the multi-scale spatial features Φ^L through weighted processing. This process is encapsulated as follows:

$$\Phi^{L'} = \Phi^L \otimes \sigma(\mathcal{F}_{1\times1}^{d\to(1-\alpha)C}(\mathcal{F}_{1\times1}^{(1-\alpha)C\to d}(\Phi^L))), \quad (7)$$

where d is set as $(1-\alpha)C/4$.

Complementary Strengths Strategy. The function of this strategy is to recouple the complementary features and achieve efficient inter-frequency communication. we propose a parameter-free manner to add the low/high-frequency features in cross-modality image to another features, where:

$$X_I^{H'} \Leftarrow X_I^{H'} + X_V^{H'}, \tag{8}$$

$$X_V^{L'} \Leftarrow X_V^{L'} + X_I^{L'}. \tag{9}$$

For each of frequency features witthin one modality, we concatenate both and use a 3×3 convolution layer $\mathcal{F}(\cdot)$ to obtain the enhanced modality-shared features. The final output is formulated as follows:

$$Y_I^S = \mathcal{F}([X_I^{H'}, X_I^{H'}]), \tag{10}$$

$$Y_V^S = \mathcal{F}([X_V^{H'}, X_V^{H'}]). \tag{11}$$

Fig. 2 shows a detailed illustration of the LFU, HFU and fusion stragy, where we intuitively demonstrate how they work by synergistically capturing high/low-frequency spatial information.

Multimodal Reconstruction Mechanism

As mentioned above, the Feature Decomposition Encoder focuses on explicitly extracting valuable frequency information. To take full advantage of complementary information, we further integrate a Multimodal Reconstruction Mechanism into our FD^2-Net. It aims to learn the discriminative and complementary features of each modality while augmenting the overall representation capability. As showcased in Fig. 2, this mechanism has two components: feature-level cross-maske and Cross-Reconstruction Unit (CRU).

Feature-Level Complementary Mask. To better utilize the multimodal information, avoid the network always learning from a single image. We design an efficient feature augmentation strategy to train FD^2-Net. As shown in Fig. 2, we perform an asymmetric mask of local information, which denoted as:

$$\mathrm{M}_{all} = \mathrm{M}_I \bigcup \mathrm{M}_V,\ \mathrm{M}_I \mid \mathrm{M}_V = 1, \tag{12}$$

where M_I and M_V represent the infrared mask and visible mask, respectively. M_{all} represents the total unseen area, accounting for 30% of the feature map. Such a design allows the network to only obtain valid information from the position corresponding to the opposite modality of the masked area.

Multimodal Image Reconstruction. As mentioned in the introduction, the information loss caused by feature extraction leads to difficulties for the detector to localize and identify objects. To address the challenge, we introduce Cross-Reconstruction Unit (CRU) to learn the complementary features through fine-grained local and coarse-grained global interactions. Note that CRU is a generic image reconstruction network, and we only take the visible image as an example to explain the working of CRU. The process can be expressed as follows (where the Rectified Linear Unit (ReLU) is omitted for brevity):

$$x_v = Conv_{3\times3}(x_v), \tag{13}$$

$$x'_v = \mathrm{CA}(x_v, x_i) + \mathcal{F}_E(Conv_{3\times3}(\mathcal{F}_S([x_v, x_i]))), \tag{14}$$

$$f_v = Conv_{1\times1}(Conv_{1\times1}(TransConv(x'_v))), \tag{15}$$

where $\mathrm{CA}(\cdot)$ represents the cross-attention layer. $\mathcal{F}_S$ and $\mathcal{F}_E$ are feature squeeze and excitation operations, same as (Hu, Shen, and Sun 2018). For the infrared and visible image, the outputs of CRU are f_i and f_v.

Training Loss

The total loss function comprises the image reconstruction loss $\mathcal{L}_{rc}$ and the detection loss $\mathcal{L}_{det}$. The reconstruction loss is computed using the mean squared error (MSE) loss between the original and reconstructed images, which is formulated as follows:

$$\mathcal{L}_{rc} = 1/2 \left\| f_i - I \right\|_2 + 1/2 \left\| f_v - V \right\|_2, \tag{16}$$

where f_i and f_v are the reconstructed infrared and visible features, respectively. I and V denote the input infrared and visible images, respectively. The detection loss, consistent with the previous algorithm, comprises classification loss $\mathcal{L}_{cls}$, localization loss $\mathcal{L}_{box}$, and confidence loss $\mathcal{L}_{obj}$:

$$\mathcal{L}_{det} = \mathcal{L}_{cls} + \mathcal{L}_{box} + \mathcal{L}_{obj}. \tag{17}$$

The overall loss function is defined as follows:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{rc} + \lambda_2 \mathcal{L}_{det}. \tag{18}$$

The λ_1 and λ_2 are the hyperparameters to balance the two losses during training.

Experiments

Experimental Settings

Datasets. The proposed model is evaluated against SOTA methods using three IVOD benchmark datasets: (1) **LLVIP** dataset (Jia et al. 2021) is a prominent large-scale pedestrian dataset specifically collected in low-light conditions, predominantly showcasing extremely dark scenes. It ensures meticulous spatial and temporal alignment between all infrared and visible image pairs, concentrating solely on pedestrian detection. (2) **FLIR** dataset offers a highly challenging multispectral object detection benchmark, encompassing both day and night scenes. In this study, we utilized the "aligned" version (Zhang et al. 2020). It comprises 5,142 precisely aligned infrared-visible image pairs, with 4,129 pairs allocated for training and 1,013 pairs reserved for testing. The dataset encompasses three primary object categories: People, Cars, and Bicycles. (3) **M^{3}FD** dataset (Liu et al. 2022a) comprises 4,200 pairs of RGB and thermal images. It includes six categories of objects: People, Cars, Buses, Motorcycles, Lamps, and Trucks. Following prior work (Zhao et al. 2023b), we employ a random splitting method to delineate the training and validation sets. Specifically, 80% of the images are allocated to the training set, with the remaining images assigned to the validation set.

Methods	↑F1	↑Precision	↑Recall	↑mAP_{50}	↑mAP_{75}
Infrared	82.6	89.8	76.5	87.6	45.9
Visible	83.5	90.7	77.4	88.5	46.3
DensFuse	88.8	91.5	86.4	89.4	58.2
SDNet	88.8	90.5	87.2	90.8	63.1
U2Fusion	89.4	90.5	88.3	91.2	61.5
CDDFuse	90.9	90.5	**91.3**	93.6	65.7
MetaF	88.6	91.1	86.3	92.7	65.5
LRRNet	91.4	93.1	89.9	94.8	68.8
SegMiF	91.3	<u>93.5</u>	89.2	94.3	67.1
TarDAL	89.9	92.3	87.6	93.3	62.4
DDFM	90.9	93.0	88.9	94.1	64.6
CSSA	89.3	91.6	87.5	92.7	65.3
TFDet	<u>91.5</u>	92.5	<u>90.4</u>	<u>95.4</u>	<u>68.9</u>
Ours	**91.7**	**94.2**	89.4	**96.2**	**70.0**

Table 1: Comparison of FD^{2}Net and SOTA methods on **LLVIP** dataset. The best and second best performance are highlighted in **bold** and <u>underline</u>.

Methods	People	Car	Bicycle	↑mAP_{50}	↑mAP_{75}
Infrared	77.2	85.2	57.9	73.7	34.0
Visible	65.6	73.8	48.7	62.7	25.9
DensFuse	78.7	85.8	61.4	75.3	35.0
SDNet	81.0	87.3	64.2	77.5	33.1
U2Fusion	82.7	87.8	67.7	79.4	36.5
CDDFuse	82.3	87.2	<u>72.9</u>	80.8	39.4
MetaF	83.3	<u>89.2</u>	71.1	81.4	40.7
LRRNet	83.3	88.8	69.7	80.6	41.0
SegMiF	**85.3**	86.9	72.8	81.5	40.9
TarDAL	85.1	85.3	69.3	79.9	37.9
DDFM	84.5	87.9	71.5	81.2	40.2
CSSA	83.2	86.7	68.6	79.4	37.2
TFDet	<u>85.2</u>	87.5	71.9	<u>81.7</u>	<u>41.3</u>
Ours	**85.3**	**89.9**	**73.2**	**82.9**	**42.5**

Table 2: Comparison of FD^{2}Net and SOTA methods on **FLIR** dataset. The best and second best performance are highlighted in **bold** and <u>underline</u>.

Methods	Peo	Car	Bus	Mot	Lam	Tru	↑mAP_{50}
Infrared	80.6	88.7	78.6	63.7	69.9	66.2	74.6
Visible	69.4	90.6	78.7	69.3	86.2	71.4	77.6
DensFuse	76.3	91.8	79.3	72.7	77.0	72.5	78.4
SDNet	79.5	92.6	81.0	67.1	84.2	69.4	79.0
U2Fusion	77.3	91.3	81.1	73.0	85.1	<u>72.8</u>	80.1
CDDFuse	81.1	93.2	<u>82.3</u>	74.0	87.7	72.7	81.9
MetaF	81.6	93.3	81.9	74.8	87.3	70.8	81.6
LRRNet	79.7	92.0	80.4	73.6	86.5	68.8	80.2
SegMiF	<u>82.4</u>	<u>93.4</u>	81.8	<u>75.7</u>	86.7	71.1	<u>82.2</u>
TarDAL	81.0	93.2	81.5	71.2	87.0	68.2	80.6
DDFM	82.0	93.1	82.2	73.6	**87.9**	71.0	81.7
Ours	**83.7**	**93.6**	**82.7**	**78.1**	<u>87.8</u>	**73.8**	**83.5**

Table 3: Comparison of FD^{2}Net and SOTA methods on **M^{3}FD** dataset. The best and second best performance are highlighted in **bold** and <u>underline</u>.

Implementation Details. To ensure fairness, we follow the same dataset processing approach as other mainstream methods (Fu et al. 2023a). FD^{2}Net is built upon the SOTA detector YOLOv5 (Jocher 2020). For evaluation, we report F1-Score, Precision, Recall, and Average Precision, consistent with prior research. Xavier initialization (Glorot and Bengio 2010) is used to initialize parameters, and the model is trained for 150 epochs using SGD (Robbins and Monro 1951) with an initial learning rate of 0.01, weight decay of 10^{-4}, and momentum of 0.9.

Main Results

We compare our proposed FD^{2}Net with several baseline and SOTA methods, including SDNet (Zhang and Ma 2021), TarDAL (Liu et al. 2022a), DensFuse (Li and Wu 2018), U2Fusion (Xu et al. 2020), CDDFuse (Zhao et al. 2023b), SegMiF (Liu et al. 2023), DDFM (Zhao et al. 2023c), MetaF (Zhao et al. 2023a), LRRNet (Li et al. 2023a), CSSA (Cao et al. 2023), and TFDet (Zhang et al. 2024). These methods are built on the YOLOv5 detector to measure their detection performance.

Comparison Results on LLVIP. The results presented in Table 1 demonstrate that our method effectively fuses similar and complementary features in infrared and visible images, significantly enhancing the network's representational capability. Compared to single-modality methods, FD^{2}Net outperforms both Infrared and Visible, with substantial improvements of **8.6%** and **7.7%**, respectively. Furthermore, when compared to other SOTA networks, FD^{2}Net consistently surpasses them, showing an improvement in mAP_{50} by **1.4%-6.8%**. These results indicate that our proposed method markedly enhances IVOD tasks performance.

Comparison Results on FLIR. As illustrated in Table 2, FD^{2}Net demonstrates exceptional performance, establishing new SOTA benchmarks for mAP_{50} and mAP_{75} at **82.9%** and **41.9%**, respectively. Specifically, our method surpasses CDDFuse and SegMiF by **+2.1%** and **+1.4%** in terms of mAP_{50}. When the threshold is increased to 0.75, the miss rate for other methods rises more significantly than for FD^{2}Net, indicating our method's superior detection accuracy. For instance, it achieves **42.5%** mAP_{75}, improving by **1.5%** over the previous best model, LRRNet.

Comparison Results on M^{3}FD. The comparative results on the M^{3}FD dataset are summarized in Table 3. Our proposed method achieves a mAP_{50} of **83.5%**, establishing new records. In addition, we present the detection accuracy for each category. Notably, in the "People" and "Motorcycle" categories, FD^{2}Net achieves improvements of **1.3%** and **2.4%** over the previous best method. This suggests that our method possesses a superior ability to detect weak and small objects.

Visual Comparisons. The qualitative results are depicted in Fig. 4. The green boxes are detection results, while the red dashed boxes mark missed objects (false negatives). It is evident that the predictions made by previous methods suffer from missed detections, especially for small and occluded objects in images. FD^{2}Net effectively captures robust shared and discriminative specific information related to detected objects, resulting in superior performance across various challenging scenarios.

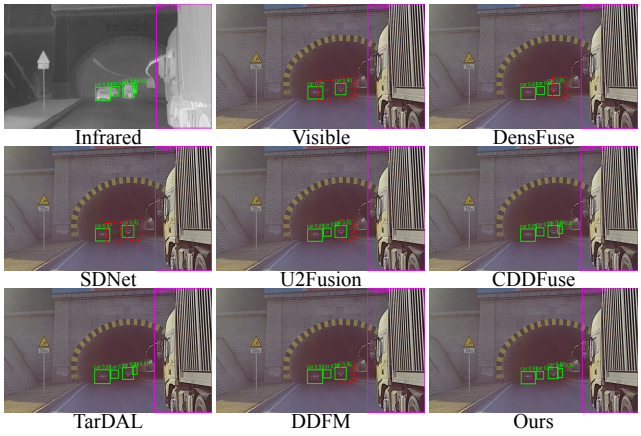

Figure 3: Visual comparison of FD^{2}Net with 10 SOTA methods. Green boxes are detection results, while red dashed boxes mark missed objects (false negatives).

	Freq-Dec Encoder		Cross-Rec Module		$\uparrow mAP_{50}$	$\uparrow mAP_{75}$
	HFU	LFU	CRU	Mask		
I	-	-	-	-	90.9	62.7
II	✓	-	-	-	92.4	64.4
III	✓	✓	-	-	94.7	66.3
IV	✓	✓	✓	-	95.6	69.6
Ours	✓	✓	✓	✓	**96.2**	**70.0**

Table 4: Ablation study of FD^{2}Net components. HFU: High-Frequency Unit, LFU: Low-Frequency Unit, CRU: Cross-Reconstruction Unit. Mask: Complementary Mask Strategy.

Ablation Study

In this section, we present the ablation study results on the LLVIP dataset to evaluate the relative effectiveness of different components in FD^{2}Net.

Architecture of FD^{2}Net. Compared to the baseline (Exp.I), the introduction of HFU (Exp.II) and LFU (Exp.III) for enhanced feature extraction improves mAP_{50} by **1.5%** and **2.3%**, respectively. Incorporating the multimodal image reconstruction strategy CRU (Exp.IV) into FD^{2}Net results in a **0.9%** improvement in mAP_{50}. Notably, mAP_{75} exhibits a substantial improvement of **3.3%**, indicating that object position perception can be significantly enhanced through image reconstruction. The feature representation capability can be further enhanced by employing an asymmetric feature mask, leading to increases of **2.6%** in AP_{50} and **1.8%** in AP_{75}. These ablation results show the effectiveness of the major components in the proposed method.

Effect of HFU and LFU. Our Feature Decomposition Encoder (FDE) comprises two components: high-frequency attention (HFU) and low-frequency refinement (LFU). To evaluate their effectiveness, we replaced the C2f blocks in YOLOv5n with either HFU or LFU. As shown in Table 5, using HFU (YOLOv5n+H) or LFU (YOLOv5n+L) alone resulted in performance drops of 3.2% and 2.8%, respectively, indicating that neither component alone effectively captures the complementary features of infrared-visible images. We further explored three integration strategies: sequential high-to-low (H+L), sequential low-to-high (L+H), and parallel (H&L). The parallel combination achieved the best performance, significantly improving mAP_{50} of **94.7%**, with reduced parameters and FLOPs. Thus, we adopt the parallel (H&L) design for FDE to maximize model performance.

	Description	↓Params.	↓FLOPs	$\uparrow mAP_{50}$
I	YOLOv5n	3.01 M	8.1 G	90.9 %
II	YOLOv5n + H	**2.72 M**	**7.6 G**	91.5 %
III	YOLOv5n + L	2.73 M	7.7 G	91.9 %
IV	YOLOv5n + H + L	2.77 M	8.0 G	92.8 %
V	YOLOv5n + L + H	2.77 M	8.0 G	92.6 %
VI	YOLOv5n + H & L	2.75 M	7.8 G	**94.7** %

Table 5: Expermental results with different combination methods of LFU and HFU on LLVIP dataset.

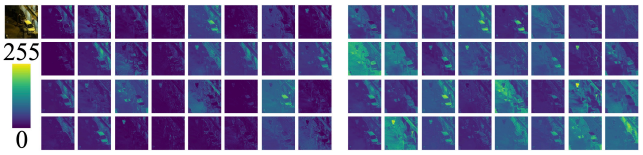

Figure 4: Left: Features from the original YOLOv5n, Right: Features from the proposed FD^{2}Net.

Feature maps visualization. To investigate the feature representation capabilities of the proposed FD^{2}Net, we visualize the feature maps from the second stage of both the original YOLOv5 and FD^{2}Net. As illustrated in Fig. 4, the feature patterns produced by FD^{2}Net are significantly enriched compared to the original YOLOv5. This approach not only reduces redundant features but also strengthens and diversifies representative features.

Conclusion

In this paper, we introduce a Frequency-Driven Feature Decomposition Network (FD^{2}Net) specifically designed for infrared-visible object detection tasks. It efficiently models high-frequency and low-frequency features, thereby facilitating the extraction of valuable complementary information. Furthermore, aided by the multimodal reconstruction mechanism, the complementary information within multimodal images is more effectively exploited. Extensive qualitative and quantitative experiments demonstrate that the proposed network attains state-of-the-art performance across competitive infrared-visible object detection benchmarks.

Acknowledgments

This work was supported in part by the National Science and Technology Major Project under Grant 2022ZD0117103, in part by the National Natural Science Foundation of China under Grants 62072354 and 62172222, in part by the Fundamental Research Funds for the Central Universities under Grants QTZX23084 and XJSJ24015, in part by the Natural Science Basic Research Program of Shaanxi under Grants 2024JC-YBQN-0732 and 2024JC-YBQN-0340, and in part by a grant from the Innovation Capability Support Program of Shaanxi under Grant 2023-CX-TD-08.

References

Bao, X.; Qin, J.; Sun, S.; Wang, X.; and Zheng, Y. 2024. Relevant Intrinsic Feature Enhancement Network for Few-Shot Semantic Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 765–773.

Behera, A.; Wharton, Z.; Hewage, P. R.; and Bera, A. 2021. Context-aware attentional pooling (cap) for fine-grained visual classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 929–937.

Cai, Z.; and Vasconcelos, N. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6154–6162.

Cao, Y.; Bin, J.; Hamari, J.; Blasch, E.; and Liu, Z. 2023. Multimodal object detection by channel switching and spatial attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 403–411.

Ding, X.; Guo, Y.; Ding, G.; and Han, J. 2019. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1911–1920.

Ding, X.; Zhang, X.; Han, J.; and Ding, G. 2022. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11963–11975.

Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; and Sun, J. 2021. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13733–13742.

F.A. Group. 2018. Flir thermal dataset for algorithm training.

Fu, H.; Wang, S.; Duan, P.; Xiao, C.; Dian, R.; Li, S.; and Li, Z. 2023a. Lraf-net: Long-range attention fusion network for visible–infrared object detection. *IEEE Transactions on Neural Networks and Learning Systems*.

Fu, H.; Wang, S.; Duan, P.; Xiao, C.; Dian, R.; Li, S.; and Li, Z. 2023b. LRAF-Net: Long-Range Attention Fusion Network for Visible–Infrared Object Detection. *IEEE Transactions on Neural Networks and Learning Systems*, 1–14.

Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 249–256.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141.

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, 448–456.

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

Jiang, B.; Luo, R.; Mao, J.; Xiao, T.; and Jiang, Y. 2018. Acquisition of localization confidence for accurate object detection. In *Proceedings of the European Conference on Computer Vision*, 784–799.

Jocher, G. 2020. YOLOv5 by Ultralytics.

Li, H.; and Wu, X.-J. 2018. DenseFuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5): 2614–2623.

Li, H.; Xu, T.; Wu, X.-J.; Lu, J.; and Kittler, J. 2023a. Lrrnet: A novel representation learning guided fusion network for infrared and visible images. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 45(9): 11040–11052.

Li, K.; Wang, D.; Hu, Z.; Zhu, W.; Li, S.; and Wang, Q. 2024. Unleashing Channel Potential: Space-Frequency Selection Convolution for SAR Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17323–17332.

Li, Y.; Hou, Q.; Zheng, Z.; Cheng, M.-M.; Yang, J.; and Li, X. 2023b. Large selective kernel network for remote sensing object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16794–16805.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2980–2988.

Liu, J.; Fan, X.; Huang, Z.; Wu, G.; Liu, R.; Zhong, W.; and Luo, Z. 2022a. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5802–5811.

Liu, J.; Liu, Z.; Wu, G.; Ma, L.; Liu, R.; Zhong, W.; Luo, Z.; and Fan, X. 2023. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8115–8124.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision*, 21–37.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022b. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11976–11986.

Nair, V.; and Hinton, G. E. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the International Conference on Machine Learning*, 807–814.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28.

Robbins, H.; and Monro, S. 1951. A stochastic approximation method. *The Annals of Mathematical Statistics*, 400–407.

Singh, B.; Najibi, M.; and Davis, L. S. 2018. Sniper: Efficient multi-scale training. *Advances in Neural Information Processing Systems*, 31.

Xu, H.; Ma, J.; Jiang, J.; Guo, X.; and Ling, H. 2020. U2Fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1): 502–518.

Zhang, H.; Fromont, E.; Lefevre, S.; and Avignon, B. 2020. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *2020 IEEE International Conference on Image Processing*, 276–280.

Zhang, H.; and Ma, J. 2021. SDNet: A versatile squeeze-and-decomposition network for real-time image fusion. *International Journal of Computer Vision*, 129(10): 2761–2785.

Zhang, J.; Lei, J.; Xie, W.; Fang, Z.; Li, Y.; and Du, Q. 2023a. SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–15.

Zhang, L.; Zhu, X.; Chen, X.; Yang, X.; Lei, Z.; and Liu, Z. 2019. Weakly Aligned Cross-Modal Learning for Multispectral Pedestrian Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Zhang, R.; Li, L.; Zhang, Q.; Zhang, J.; Xu, L.; Zhang, B.; and Wang, B. 2023b. Differential Feature Awareness Network within Antagonistic Learning for Infrared-Visible Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 1–1.

Zhang, X.; Zhai, H.; Liu, J.; Wang, Z.; and Sun, H. 2023c. Real-time infrared and visible image fusion network using adaptive pixel weighting strategy. *Information Fusion*, 99: 101863.

Zhang, X.; Zhang, X.; Wang, J.; Ying, J.; Sheng, Z.; Yu, H.; Li, C.; and Shen, H.-L. 2024. Tfdet: Target-aware fusion for rgb-t pedestrian detection. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhang, Y.; Yu, H.; He, Y.; Wang, X.; and Yang, W. 2023d. Illumination-guided RGBT object detection with inter-and intra-modality fusion. *IEEE Transactions on Instrumentation and Measurement*, 72: 1–13.

Zhao, W.; Xie, S.; Zhao, F.; He, Y.; and Lu, H. 2023a. Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13955–13965.

Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Xu, S.; Lin, Z.; Timofte, R.; and Van Gool, L. 2023b. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5906–5916.

Zhao, Z.; Bai, H.; Zhu, Y.; Zhang, J.; Xu, S.; Zhang, Y.; Zhang, K.; Meng, D.; Timofte, R.; and Van Gool, L. 2023c. DDFM: denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8082–8093.

Zhou, K.; Chen, L.; and Cao, X. 2020. Improving multispectral pedestrian detection by addressing modality imbalance problems. In *Proceedings of the European Conference on Computer Vision*, 787–803.