

A Dual-Stream Network with Spatial Alignment and Swin Transformer Fusion for Visible-Infrared Object Detection

Jingtao Hu

Soochow University

Yunxiang Zhu

Soochow University

Yangkun Zhu

Soochow University

Zhenzhou Shao

Capital Normal University

Donghan Kim

Kyung Hee University

Yibo Zou

`yibo.zou@suda.edu.cn`

Soochow University

Junying Yang

Soochow University

Research Article

Keywords:

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11115500/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Dual-Stream Network with Spatial Alignment and Swin Transformer Fusion for Visible-Infrared Object Detection

Jingtao Hu¹, Yunxiang Zhu¹, Yangkun Zhu¹, Zhenzhou Shao²,
Donghan Kim³, Junying Yang^{1*}, Yibo Zou^{1*}

¹School of Optoelectronic Science and Technology, Soochow University,
Suzhou, 215006, China.

²College of Information Engineering, Capital Normal University,
Beijing, China.

³Department of Electronic Engineering, Kyung Hee University,
Yongin-si, 17104, Gyeonggi-do, Republic of Korea.

*Corresponding author(s). E-mail(s): jyyang22@suda.edu.cn;
yibo.zou@suda.edu.cn;

Abstract

Visible-infrared object detection has attracted increasing attention due to the complementary characteristics of visible and infrared modalities under complex imaging conditions. Nevertheless, effective cross-modal feature fusion remains challenging because visible and infrared images usually suffer from spatial misalignment, modality-specific appearance discrepancy, and redundant feature interaction. In this paper, we propose a visible-infrared object detection network based on dual-stream feature extraction, spatial alignment, and Swin Transformer-based cross-modal feature fusion. Specifically, a dual-stream backbone is first constructed to extract modality-specific representations from visible and infrared images. Inspired by existing wavelet-based feature modeling strategies, wavelet transform is adopted as an auxiliary frequency-domain operation in the backbone and neck to preserve structural and detail information during feature extraction and aggregation. To alleviate feature-level spatial misalignment, a spatial alignment module, termed OffsetNet, is introduced to predict task-oriented sampling offsets and align visible features to the infrared reference feature space. After spatial alignment, a Swin Transformer-based cross-modal feature fusion module, termed SwinFusion, is designed to perform intra-modal feature modeling and inter-modal feature fusion, enabling effective exploitation

of complementary information between the two modalities. Experimental results on public visible-infrared datasets show that the proposed method consistently improves detection accuracy over the baseline. Comprehensive ablation studies further demonstrate the effectiveness of OffsetNet and SwinFusion.

Keywords: Visible-infrared object detection, Dual-stream network, Spatial alignment, Cross-modal feature fusion, Swin Transformer

1 Introduction

Object detection has been widely studied and applied in many real-world scenarios, such as autonomous driving, remote sensing, intelligent surveillance, and unmanned aerial perception [1–3]. Although visible images can provide rich color, texture, and structural details, their imaging quality is easily affected by illumination changes, shadows, haze, and nighttime environments. In contrast, infrared images are less sensitive to visible illumination and can highlight objects according to thermal radiation differences, making them more reliable under low-light or adverse imaging conditions [4]. Therefore, visible-infrared object detection has attracted increasing attention because it can exploit the complementary characteristics of the two modalities to improve detection robustness in complex scenes [5].

Despite the advantages of visible-infrared fusion, effectively integrating the two modalities remains challenging. First, visible and infrared images are usually captured by different sensors, which may have different viewpoints, focal lengths, resolutions, fields of view, and temporal synchronization errors. As a result, the paired images are often weakly aligned rather than strictly registered, leading to unreliable feature correspondence during fusion [6]. Second, visible and infrared modalities have different imaging mechanisms. Visible images mainly describe surface appearance, color, and texture, whereas infrared images mainly reflect thermal radiation and object contours. Such modality discrepancy makes direct feature addition or concatenation insufficient for extracting complementary information [7]. Third, during feature extraction and aggregation, conventional convolutional operations may cause the loss of structural and detail information, especially for small objects and object boundaries. Therefore, it is necessary to improve feature representation while maintaining effective cross-modal alignment and fusion.

Prior research has explored various approaches for visible-infrared feature fusion. Liu et al. [8] adopted a two-stream network architecture for RGB-T pedestrian detection and investigated different fusion strategies, including early fusion, halfway fusion, and late fusion. Their experimental results showed that halfway fusion achieves better performance than other fusion strategies. Although simple operations such as feature concatenation, addition, and network-in-network (NIN) fusion are efficient, they lack the ability to explicitly model fine-grained cross-modal relationships. Various attention-based methods have subsequently been proposed to achieve more adaptive feature fusion [9–12].

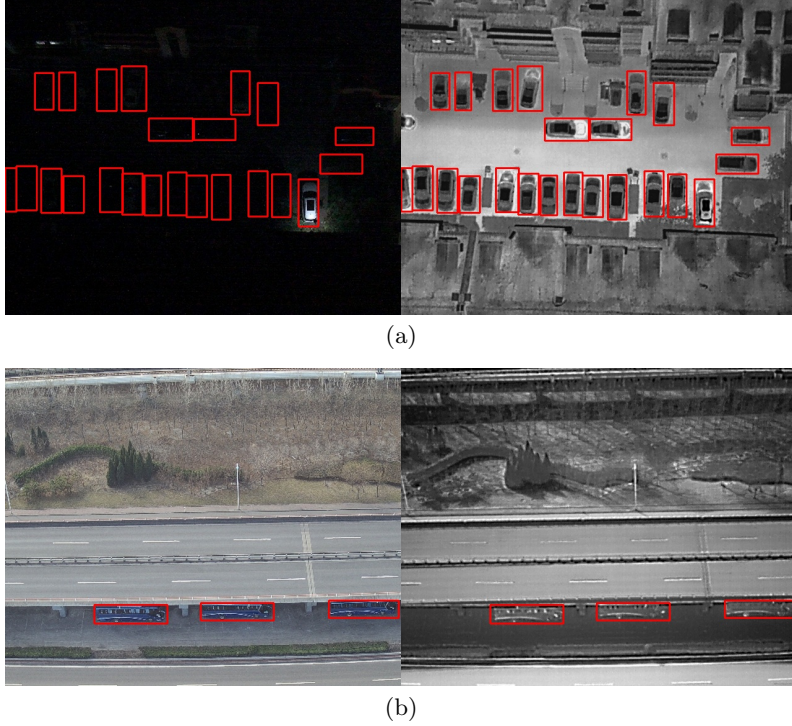


Fig. 1 Examples of visible-infrared image pairs. (a) and (b) show two representative visible-infrared image pairs. Visible images provide rich color, texture, and appearance information, while infrared images highlight thermal responses and object contours under complex imaging conditions, demonstrating the complementary characteristics of the two modalities.

Recently, attention-based and Transformer-based methods have been introduced to enhance cross-modal interaction. Cross-attention mechanisms have been widely used to capture complementary information between visible and infrared features [5, 7]. Meanwhile, alignment-aware methods have been proposed to alleviate weak alignment by learning adaptive feature correspondence or feature-level offsets [6, 13, 14]. In addition, frequency-domain modeling based on wavelet transform has shown the potential to preserve structural and detail information for multimodal representation learning [15–17]. However, robust visible-infrared detection still requires a unified framework that can simultaneously consider modality-specific feature extraction, feature-level spatial alignment, and effective cross-modal feature fusion.

To address the above issues, this paper proposes a visible-infrared object detection network based on dual-stream feature extraction, spatial alignment, and Swin Transformer-based cross-modal feature fusion. The proposed framework first constructs a dual-stream backbone to extract multi-scale features from visible and infrared images, respectively. In addition, inspired by WaveMamba [17], wavelet-based feature modeling is adopted as an auxiliary frequency-domain operation in the backbone and neck to preserve structural and detail information during feature extraction and aggregation.

For cross-modal interaction, a spatial alignment module, termed OffsetNet, is introduced before feature fusion. OffsetNet predicts task-oriented sampling offsets to align visible features to the infrared reference feature space, thereby reducing local spatial discrepancies between the two modalities. It should be noted that the predicted offsets are not intended to represent a strict physical registration field between visible and infrared images. Instead, they serve as feature-level sampling offsets optimized for the detection task, which helps alleviate misalignment before cross-modal fusion [6, 13].

After spatial alignment, a Swin Transformer-based cross-modal feature fusion module, termed SwinFusion, is further designed to exploit complementary information between visible and infrared features. Motivated by previous multispectral detection studies that model intra-modality and inter-modality relationships for feature fusion [7, 9], SwinFusion adopts an Intra-Inter Swin Fusion design. Specifically, the aligned visible feature and infrared feature are first independently processed for intra-modal feature modeling, and then inter-modal feature fusion is performed to enhance cross-modal complementary representation. In this way, the network avoids directly mixing unaligned features and obtains more discriminative bimodal features for detection.

The main contributions of this paper are summarized as follows:

- A unified dual-modal detection framework is proposed for visible-infrared object detection, which integrates a feature-level spatial alignment module and a cross-modal feature fusion module into a dual-stream detection network. By performing alignment-before-fusion interaction at multiple detection scales, the framework can reduce cross-modal feature inconsistency and obtain more reliable multimodal representations.
- A spatial alignment module, termed OffsetNet, is introduced to alleviate feature-level misalignment between visible and infrared modalities. Instead of performing strict image-level registration, OffsetNet predicts task-oriented sampling offsets and aligns visible features to the infrared reference feature space before cross-modal fusion.
- A Swin Transformer-based cross-modal feature fusion module, termed SwinFusion, is designed to strengthen visible-infrared feature interaction. By adopting an Intra-Inter Swin Fusion strategy, the module first models intra-modal contextual information and then performs inter-modal feature fusion, enabling more effective exploitation of cross-modal complementary information.

2 Related Work

2.1 Visible-Infrared Object Detection

Visible-infrared object detection aims to improve detection robustness by exploiting complementary information from visible and infrared modalities. Visible images usually contain rich color, texture, and appearance details, but they are sensitive to illumination changes and may suffer from severe degradation under low illumination or adverse weather conditions. In contrast, infrared images capture thermal radiation

information and can highlight object contours under poor lighting conditions. Therefore, visible-infrared detection has been widely investigated in challenging scenarios such as nighttime perception, autonomous driving, surveillance, and remote sensing.

Early multispectral detection methods mainly extended single-modal detectors to dual-stream architectures, in which visible and infrared features are extracted separately and fused at different stages. Liu et al. [8] investigated early fusion, halfway fusion, and late fusion strategies for RGB-T pedestrian detection, showing that halfway fusion can achieve better performance than other fusion stages. Subsequent works further explored feature-level fusion because it can preserve modality-specific representations while enabling end-to-end optimization for detection. Zhang et al. [9] proposed a guided attentive feature fusion method, which introduces intra-modality and inter-modality attention to dynamically weigh and fuse multispectral features. Dasgupta et al. [10] designed a spatio-contextual deep network with graph attention and contextual refinement for multimodal pedestrian detection in autonomous driving scenarios. Fang and Wang [11] proposed a cross-modality attentive feature fusion module for multispectral remote sensing object detection, which enhances modality-specific features and selects useful modality-shared features. Li et al. [12] introduced confidence-aware fusion based on Dempster-Shafer theory to dynamically integrate multispectral information according to modality confidence. These methods demonstrate that adaptive feature fusion is effective for visible-infrared detection, but robust cross-modal interaction and spatially consistent feature fusion remain challenging.

2.2 Cross-modal Feature Fusion

Feature fusion is a key component in visible-infrared object detection. Simple fusion operations, such as feature addition, concatenation, and network-in-network (NIN) fusion, are efficient and have been widely used in early dual-stream detectors. However, these operations usually treat features from different modalities in a fixed manner and lack the ability to explicitly model fine-grained cross-modal relationships. To address this issue, attention-based fusion methods have been proposed to dynamically select or enhance useful information from different modalities.

Guided attentive feature fusion (GAFF) introduces intra-modality and inter-modality attention to adaptively weigh multispectral features under different conditions [9]. Confidence-aware fusion further estimates the reliability of each modality and fuses detection results according to modality confidence, improving the reliability of multispectral pedestrian detection [12]. Cross-modality attentive feature fusion (CMAF) decomposes modality features into common and differential components, aiming to retain modality-specific features while selecting useful shared features [11].

Recently, Transformer-based fusion methods have been introduced to strengthen global cross-modal interaction. CFT employs Transformer-based feature fusion to model long-range dependencies between RGB and thermal features [5]. ICATFusion proposes an iterative cross-attention guided fusion framework to capture complementary information from different modalities while reducing model complexity through parameter sharing [7]. JTMDet combines Transformer and Mamba structures to capture useful multispectral information while maintaining high inference efficiency [18].

EI²Det designs a cross-modal interaction module with visible-priority and infrared-priority multi-head cross-attention, and further introduces illumination-aware and edge-guided mechanisms for robust visible-infrared detection [19]. These methods show that attention-based and Transformer-based interaction can effectively enhance cross-modal feature fusion. However, directly fusing features may still introduce feature interference when visible and infrared features are locally misaligned or have large modality discrepancies.

2.3 Weakly Aligned Cross-modal Feature Alignment

Cross-modal misalignment is another important challenge in visible-infrared object detection. Since visible and infrared images are usually captured by different sensors, the paired images may suffer from inconsistent viewpoints, focal lengths, resolutions, imaging time, or field of view. Such weak alignment causes feature mismatch between corresponding regions, which may degrade fusion quality and detection performance.

AR-CNN is one of the representative methods for weakly aligned multispectral pedestrian detection. It points out that real color-thermal image pairs may suffer from position shift, and proposes a Region Feature Alignment module to adaptively align region features from two modalities [6]. For UAV-based RGB-IR object detection, OAFA further argues that strict image-level alignment is difficult and costly in real-world scenarios. Instead of enforcing strict registration, it estimates feature-level offsets and performs offset-guided deformable alignment and fusion for detection-oriented multimodal feature interaction [13]. DACFusion further emphasizes the importance of spatial correspondence between multispectral images and proposes a dual asymmetric cross-attention fusion strategy to interactively process features according to positional correspondence and modality characteristics [14].

These works show that task-driven alignment-aware feature interaction is more practical than strict physical registration for weakly aligned visible-infrared detection. Inspired by this observation, the proposed method adopts feature-level offset prediction before cross-modal fusion. Instead of performing strict image-level registration, the alignment module predicts task-oriented sampling offsets to align visible features to the infrared reference feature space, thereby reducing local spatial discrepancy before feature fusion.

2.4 Frequency-domain Feature Modeling

Frequency-domain modeling provides an effective auxiliary perspective for visual feature representation. Wavelet transform can decompose feature maps into low-frequency structural components and high-frequency detail components, making it suitable for separating global object structures from local edge and texture information. WaveMamba introduces a wavelet-driven fusion strategy for RGB-infrared object detection, demonstrating that wavelet-based frequency-domain modeling can effectively support multispectral feature representation [17]. WtCAFNet further introduces wavelet transform and cross-attention modality-adaptive fusion for multispectral object detection, where wavelet-related operations are used to enhance feature representation and reduce redundant information during cross-modal fusion [15]. These

studies indicate that wavelet-based feature modeling can help preserve structural and detail information in multimodal feature extraction and aggregation.

Based on the above observations, this work adopts the wavelet-driven feature modeling strategy from WaveMamba [17] as an auxiliary frequency-domain operation in the backbone and neck, rather than treating wavelet transform itself as the main methodological novelty. The proposed network mainly focuses on three aspects: constructing a dual-stream detection framework for visible and infrared modalities, introducing an OffsetNet module to reduce feature-level spatial misalignment, and designing a Swin Transformer-based cross-modal feature fusion module to strengthen intra-modal representation and inter-modal information interaction. In this way, the adopted wavelet operations provide frequency-aware feature representations, while the task-specific alignment and fusion modules further improve the effectiveness of visible-infrared object detection.

3 Method

This section introduces the proposed visible-infrared object detection network. Overall, the proposed method is built upon a dual-stream YOLO detection framework. To address modality discrepancy, spatial misalignment, and insufficient cross-modal interaction between visible and infrared images, we construct a unified dual-modal detection network that integrates feature extraction, wavelet-based feature modeling, feature-level spatial alignment, and cross-modal feature fusion. Specifically, a dual-stream backbone is first used to extract multi-scale features from visible and infrared images, respectively. Inspired by WaveMamba [17], the discrete wavelet transform (DWT) and inverse discrete wavelet transform (IDWT) are adopted as auxiliary frequency-domain operations in the backbone and neck to preserve structural and detail information during feature extraction and aggregation. In the cross-modal interaction stage, the four wavelet subbands at each feature scale are processed by an alignment and fusion pipeline with parameters shared within that scale. Different feature scales use separate module instances with independent parameters. The alignment module, termed OffsetNet, predicts task-oriented spatial sampling offsets to align visible subband features to the infrared reference feature space. Then, the fusion module, termed SwinFusion, performs intra-modal feature modeling and inter-modal feature fusion between the aligned visible subband features and infrared subband features. Finally, the fused wavelet subbands are fed into the neck, where low-frequency processing and cross-scale aggregation are combined with IDWT reconstruction before detection.

3.1 Overall Architecture

The overall architecture of the proposed network is shown in Fig. 2. Given a pair of visible image I^{gb} and infrared image I^{ir} , the network extracts multi-scale features from the two modalities through a dual-stream backbone. In our implementation, the visible and infrared branches are instantiated with YOLOv8 and YOLOv11, respectively. This setting is used as an implementation configuration within the dual-stream detection framework to extract modality-specific multi-scale features before cross-modal

alignment and fusion. It should be noted that this work does not regard the backbone configuration itself as an independent methodological contribution. Instead, the main focus of this paper is on feature-level spatial alignment and Swin Transformer-based cross-modal feature fusion. Therefore, all ablation studies are conducted under the same backbone configuration to fairly evaluate the effectiveness of OffsetNet and SwinFusion. Let $\mathcal{B}^{\text{rgb}}$ and $\mathcal{B}^{\text{ir}}$ denote the visible and infrared branches, respectively. The multi-scale feature extraction process can be formulated as:

$$\begin{aligned}\mathcal{F}^{\text{rgb}} &= \{F_3^{\text{rgb}}, F_4^{\text{rgb}}, F_5^{\text{rgb}}\} = \mathcal{B}^{\text{rgb}}(I^{\text{rgb}}), \\ \mathcal{F}^{\text{ir}} &= \{F_3^{\text{ir}}, F_4^{\text{ir}}, F_5^{\text{ir}}\} = \mathcal{B}^{\text{ir}}(I^{\text{ir}}).\end{aligned}\tag{1}$$

where $\mathcal{F}^{\text{rgb}}$ and $\mathcal{F}^{\text{ir}}$ denote the three-level feature sets extracted from the visible and infrared branches, respectively.

Since intermediate fusion has been widely verified to be effective in visible-infrared detection, this work also adopts an intermediate fusion strategy and mainly performs cross-modal interaction on the last three feature scales used for detection. As shown in Fig. 2 and Fig. 3, after DWT-based feature decomposition, the proposed Wavelet-Guided Alignment and Fusion (WGAF) module is applied to the wavelet subbands of visible and infrared features. Different from conventional methods that directly concatenate or add features from two modalities, the proposed method first aligns the visible subband feature to the infrared reference feature space and then performs cross-modal feature fusion.

For a selected feature scale, DWT first decomposes the visible and infrared features into four wavelet subbands:

$$\begin{aligned}\{F_{LL}^m, F_{LH}^m, F_{HL}^m, F_{HH}^m\} &= \text{DWT}(F^m), \\ m &\in \{\text{rgb}, \text{ir}\}.\end{aligned}\tag{2}$$

where F_{LL}^m denotes the low-frequency structural subband, while F_{LH}^m , F_{HL}^m , and F_{HH}^m denote the high-frequency directional subbands.

Let $s \in \{LL, LH, HL, HH\}$ denote one of the four wavelet subbands. Within a given feature scale, the visible and infrared features from each subband are processed by the same alignment and fusion pipeline with shared parameters:

$$F_s^{\text{rgb-align}} = \text{OffsetNet}(F_s^{\text{rgb}}, F_s^{\text{ir}}),\tag{3}$$

$$F_s^{\text{fuse}} = \text{SwinFusion}(F_s^{\text{rgb-align}}, F_s^{\text{ir}}).\tag{4}$$

Here, the parameters of OffsetNet and those of SwinFusion are each shared across the four wavelet subbands within the same feature scale. Separate OffsetNet and SwinFusion instances are used at different feature scales, without cross-scale parameter sharing. For notational simplicity, both the scale index and the subband index are omitted in the following single-subband derivations. Therefore, F^{rgb} and F^{ir} are used to denote an arbitrary pair of visible and infrared subband features from a selected feature scale, and the same operations are applied to all selected scales and all wavelet subbands.

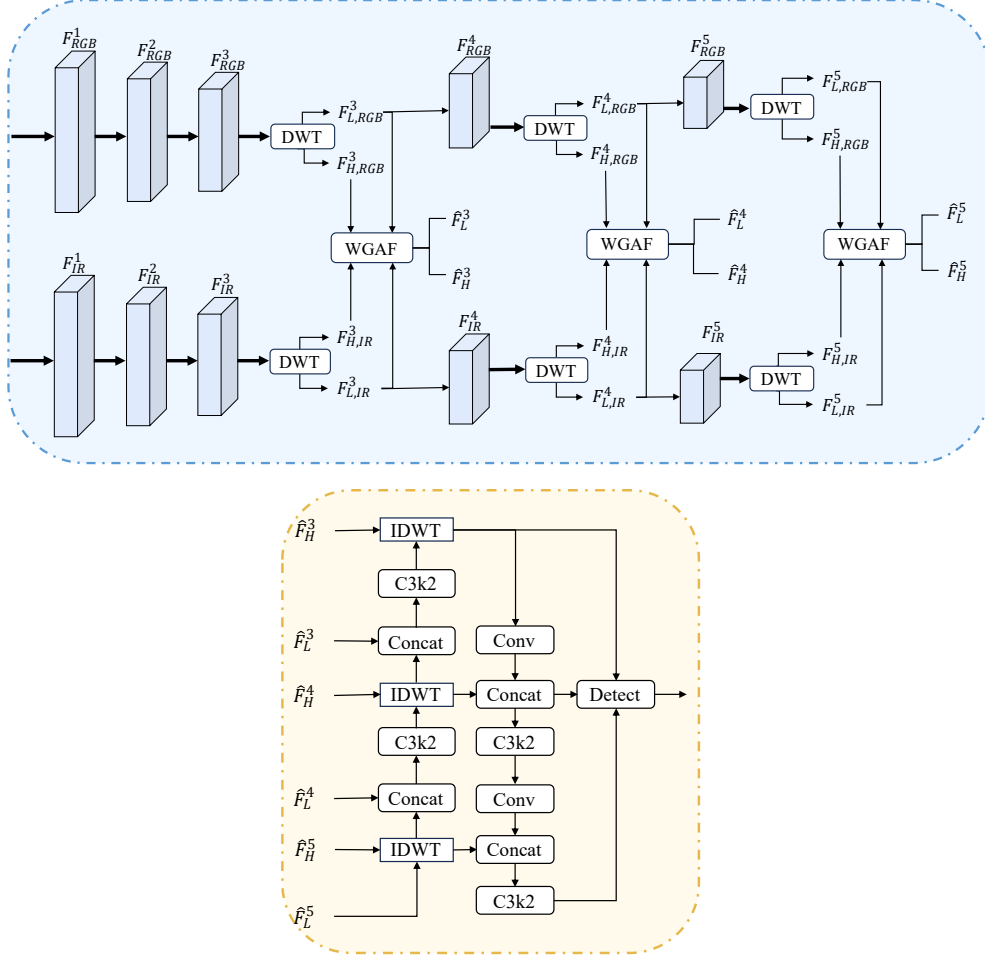


Fig. 2 Overall architecture of the proposed network with Wavelet-Guided Alignment and Fusion (WGAF). WGAF parameters are shared across subbands within each scale and are independent across scales. In the neck, low-frequency features undergo convolutional and cross-scale aggregation before IDWT reconstruction with the corresponding high-frequency features.

The fused low-frequency and high-frequency features are passed to the wavelet-based neck. Let $L_i = F_{i,LL}^{\text{fuse}}$ and $H_i = \{F_{i,LH}^{\text{fuse}}, F_{i,HL}^{\text{fuse}}, F_{i,HH}^{\text{fuse}}\}$ for $i \in \{3, 4, 5\}$. These correspond to F_L^i and F_H^i in Fig. 2. The neck combines low-frequency processing, cross-scale aggregation, and IDWT reconstruction to produce the detection features:

$$\{F_3^{\text{neck}}, F_4^{\text{neck}}, F_5^{\text{neck}}\} = \mathcal{N} \left(\{(L_i, H_i)\}_{i=3}^5 \right). \quad (5)$$

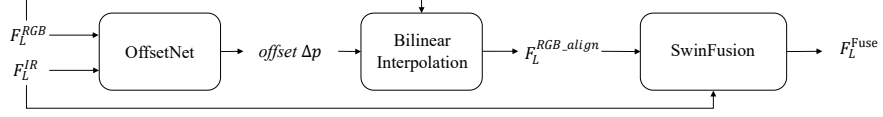


Fig. 3 Structure of WGAF for one wavelet subband. OffsetNet predicts sampling offsets for bilinear resampling of visible features, followed by SwinFusion with the infrared reference features.

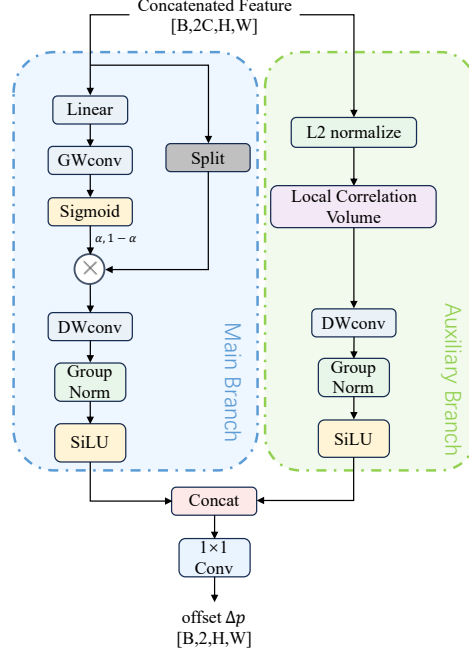


Fig. 4 Architecture of the proposed OffsetNet. OffsetNet takes the concatenated visible-infrared subband feature as input and predicts the spatial offset field through a Main Branch and an Auxiliary Branch. The Main Branch performs gated cross-modal feature weighting, while the Auxiliary Branch computes a local correlation volume within a 3×3 neighborhood to provide local matching guidance. The outputs of the two branches are concatenated and projected by a 1×1 convolution to generate the final offset field Δp .

Finally, the detection head takes the aggregated neck features and predicts object categories, bounding boxes, and confidence scores:

$$Y = \mathcal{D}(\{F_3^{\text{neck}}, F_4^{\text{neck}}, F_5^{\text{neck}}\}), \quad (6)$$

where $\mathcal{N}(\cdot)$ denotes the neck, $\mathcal{D}(\cdot)$ denotes the detection head, and Y denotes the final detection results.

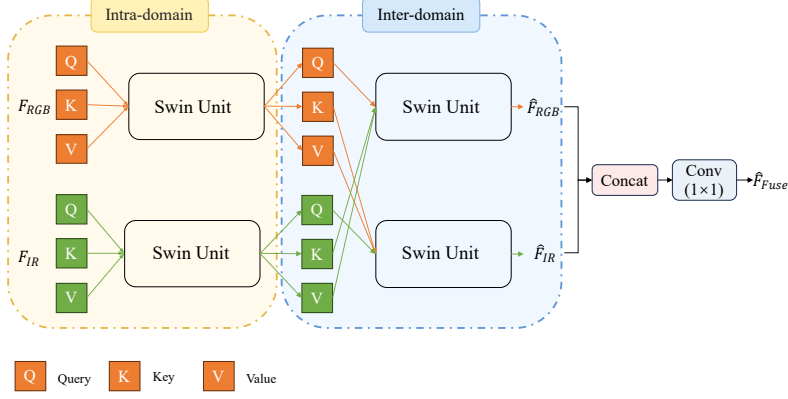


Fig. 5 Architecture of the proposed SwinFusion module. The module first performs intra-domain feature modeling for aligned visible and infrared features from the same wavelet subband, and then conducts inter-domain feature fusion to enhance cross-modal complementary representation. The enhanced features from the two modalities are concatenated and projected by a 1×1 convolution to generate the final fused subband feature.

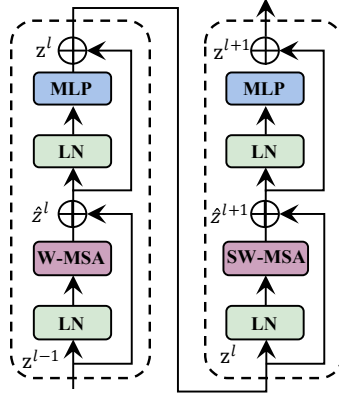


Fig. 6 Swin unit redrawn based on the Swin Transformer block [20]. The illustrated W-MSA and SW-MSA operations apply to intra-modal self-attention. In the inter-modal stage, they are replaced by regular-window and shifted-window cross-attention: queries and residual connections come from the branch being updated, while keys and values come from the complementary modality.

3.2 Dual-stream Backbone with Wavelet-based Feature Modeling

The proposed network adopts a dual-stream backbone to extract visible and infrared features separately. The visible and infrared branches generate multi-scale feature maps, which are subsequently processed by wavelet-based feature modeling operations and cross-modal alignment and fusion modules. In our implementation, the two branches are instantiated as described in the overall architecture, and the same backbone configuration is kept fixed in all ablation experiments.

Following the wavelet-driven feature modeling strategy in WaveMamba [17], DWT is introduced as an auxiliary frequency-domain operation in the backbone. For an intermediate feature F^m , where $m \in \{\text{rgb}, \text{ir}\}$, DWT decomposes it into one low-frequency subband and three high-frequency directional subbands:

$$\{F_{LL}^m, F_{LH}^m, F_{HL}^m, F_{HH}^m\} = \text{DWT}(F^m). \quad (7)$$

The LL subband mainly contains low-frequency structural information, such as object contours, semantic layout, and coarse appearance cues. The LH, HL, and HH subbands mainly contain high-frequency details, such as edges, textures, and local variations along different directions.

By adopting wavelet decomposition in the backbone, the network can explicitly access different frequency components during feature extraction, reducing the loss of structural and detail information caused by conventional convolutional downsampling. In our implementation, the four wavelet subbands are processed by the same alignment and fusion pipeline. Specifically, the LL, LH, HL, and HH subbands within each feature scale are fed into shared OffsetNet and SwinFusion modules for feature-level alignment and cross-modal fusion. The module parameters are independent across feature scales. Since the processing procedure of the four subbands is identical, the following OffsetNet and SwinFusion derivations use F^{rgb} and F^{ir} to denote an arbitrary pair of visible and infrared subband features, and both the scale index and the subband index are omitted for clarity. It should be noted that the wavelet operation is used as an adopted frequency-domain feature modeling tool in this work, while the main task-specific designs focus on feature-level spatial alignment and cross-modal feature fusion.

3.3 Spatial Alignment Module

In practical acquisition scenarios, visible and infrared images are usually captured by different sensors, and they may have differences in viewpoint, focal length, installation position, resolution, and temporal synchronization. Therefore, even after preprocessing, local misalignment may still exist at the feature level. Directly fusing unaligned features may cause inconsistency between object textures, edges, and thermal responses, thus affecting detection performance.

To mitigate this problem, a spatial alignment module, termed OffsetNet, is designed before cross-modal feature fusion. Given a visible subband feature F^{rgb} and an infrared subband feature F^{ir} from the same feature scale and wavelet subband, OffsetNet predicts the spatial sampling offsets of the visible feature with respect to the infrared reference feature space and generates the aligned visible feature $F^{\text{rgb_align}}$ through offset-guided backward sampling. Within each feature scale, the same OffsetNet instance is applied to all four wavelet subbands. Different feature scales use independent OffsetNet instances.

The two subband features are first concatenated along the channel dimension to obtain a bimodal joint feature:

$$F_{\text{cat}} = \text{Concat}(F^{\text{rgb}}, F^{\text{ir}}), \quad (8)$$

where $F_{\text{cat}} \in \mathbb{R}^{B \times 2C \times H \times W}$ denotes the concatenated bimodal subband feature.

As shown in Fig. 4, OffsetNet consists of a Main Branch and an Auxiliary Branch. The Main Branch is motivated by recent attention-based visual representation methods such as TransNeXt [21], which emphasize adaptive feature aggregation for enhancing discriminative visual responses. Based on this motivation, the Main Branch performs gated cross-modal feature weighting to adaptively emphasize informative visible and infrared feature responses for offset prediction. The Auxiliary Branch is inspired by the cost-volume-based matching mechanism in PWC-Net [22], where local feature similarities are used to estimate dense correspondences. Different from optical flow estimation, the Auxiliary Branch in OffsetNet only computes a local correlation volume to provide explicit local matching cues for detection-oriented feature alignment. The outputs of the two branches are concatenated and projected by a 1×1 convolution to generate the final spatial offset field.

In the Main Branch, the concatenated feature is first processed by a linear mapping layer and a 3×3 group-wise convolution. Then, a Sigmoid function is used to generate a spatially adaptive gating weight:

$$\alpha = \sigma(\text{GWConv}(\text{Linear}(F_{\text{cat}}))), \quad (9)$$

where $\alpha \in \mathbb{R}^{B \times C \times H \times W}$, and $\sigma(\cdot)$ denotes the Sigmoid function. Since the input feature has $2C$ channels, the group-wise convolution can model local cross-modal channel relationships with low computational cost.

Meanwhile, the concatenated feature is split into visible and infrared feature parts:

$$[X^{\text{rgb}}, X^{\text{ir}}] = \text{Split}(F_{\text{cat}}). \quad (10)$$

Then, the gating weight and its complementary weight are used to adaptively combine visible and infrared features:

$$X_{\text{fuse}} = \alpha \odot X^{\text{rgb}} + (1 - \alpha) \odot X^{\text{ir}}, \quad (11)$$

where $\odot$ denotes element-wise multiplication. This gated weighting strategy enables OffsetNet to adaptively emphasize informative modality responses at different spatial positions for offset estimation.

The weighted feature is further processed by depth-wise convolution, group normalization, and SiLU activation to generate the output of the Main Branch:

$$M = \text{SiLU}(\text{GN}(\text{DWConv}(X_{\text{fuse}}))), \quad (12)$$

where M denotes the output feature of the Main Branch.

In the Auxiliary Branch, a local correlation volume is introduced to provide local spatial correspondence information between visible and infrared subband features. Different from the Main Branch, which learns adaptive feature weighting, the Auxiliary Branch explicitly compares the infrared reference feature with neighboring visible source features within a local search region.

The visible and infrared subband features are first normalized along the channel dimension:

$$\bar{F}^{\text{rgb}} = \text{L2Norm}(F^{\text{rgb}}), \quad \bar{F}^{\text{ir}} = \text{L2Norm}(F^{\text{ir}}), \quad (13)$$

where $\text{L2Norm}(\cdot)$ denotes channel-wise L2 normalization.

For each spatial position $p = (x, y)$ in the infrared reference feature map, the local correlation volume measures the similarity between the infrared feature at position p and the visible features within a local neighborhood around p . In this work, a 3×3 local neighborhood is adopted. Therefore, the local displacement set is defined as:

$$\mathcal{D}_1 = \{(u, v) \mid u, v \in \{-1, 0, 1\}\}, \quad (14)$$

which contains 9 local displacement candidates.

For each displacement $d = (u, v) \in \mathcal{D}_1$, the local correlation response is computed by the channel-wise inner product between the normalized infrared feature at position p and the normalized visible feature at the shifted position $p + d$:

$$C(p, d) = \sum_{c=1}^C \bar{F}_c^{\text{ir}}(p) \cdot \bar{F}_c^{\text{rgb}}(p + d). \quad (15)$$

Since both features are L2-normalized along the channel dimension, this inner product can be regarded as the local cosine similarity between the infrared reference feature and the neighboring visible source feature.

By enumerating all displacement candidates in the 3×3 neighborhood, the correlation responses are stacked along the channel dimension to form the local correlation volume:

$$\mathcal{C} = \text{Concat}(\{C(\cdot, d) \mid d \in \mathcal{D}_1\}), \quad (16)$$

where $\mathcal{C} \in \mathbb{R}^{B \times 9 \times H \times W}$ denotes the local correlation volume. It encodes the local matching relationship between the infrared reference feature and the visible source feature under nine neighboring displacement candidates.

The local correlation volume is further processed by depth-wise convolution, group normalization, and SiLU activation to generate the auxiliary correlation-guided feature:

$$A = \text{SiLU}(\text{GN}(\text{DWConv}(\mathcal{C}))), \quad (17)$$

where A denotes the output feature of the Auxiliary Branch. This auxiliary feature provides explicit local matching evidence for offset prediction, helping OffsetNet estimate more stable task-oriented sampling offsets under cross-modal spatial discrepancy.

Finally, the outputs of the Main Branch and the Auxiliary Branch are concatenated and projected by a 1×1 convolution to generate the spatial offset field:

$$\Delta p = \text{Conv}_{1 \times 1}(\text{Concat}(M, A)), \quad (18)$$

where $\Delta p \in \mathbb{R}^{B \times 2 \times H \times W}$ denotes the predicted spatial offset field. For each spatial position $p = (x, y)$ in the feature map, $\Delta p(p) = (\Delta x, \Delta y)$ represents the sampling offset in the horizontal and vertical directions.

After obtaining the spatial offset field Δp , the infrared subband feature F^{ir} is used as the reference coordinate space, and offset-guided backward sampling is performed on the visible subband feature F^{rgb} to obtain the aligned visible feature $F^{\text{rgb-align}}$. Specifically, for an arbitrary spatial position $p = (x, y)$ in the infrared reference feature map, the aligned visible feature does not directly take the response from the same position in the original visible feature map. Instead, it samples the corresponding response from the original visible feature map at the offset position $p + \Delta p(p)$:

$$F^{\text{rgb-align}}(p) = F^{\text{rgb}}(p + \Delta p(p)). \quad (19)$$

In coordinate form, this can be written as:

$$F^{\text{rgb-align}}(x, y) = F^{\text{rgb}}(x + \Delta x, y + \Delta y). \quad (20)$$

Since the sampling position $p + \Delta p(p)$ is usually a continuous coordinate rather than an integer grid point, bilinear interpolation is adopted to obtain the corresponding feature response from the original visible feature map:

$$F^{\text{rgb-align}}(p) = \sum_{q \in \mathcal{N}(p + \Delta p(p))} G(q, p + \Delta p(p)) F^{\text{rgb}}(q), \quad (21)$$

where $\mathcal{N}(p + \Delta p(p))$ denotes the set of neighboring grid points around the sampling position, and $G(q, p + \Delta p(p))$ denotes the bilinear interpolation weight.

This alignment process can be regarded as a feature-level resampling strategy using infrared features as the reference. For each position in the infrared feature map, the network adaptively extracts feature responses from the corresponding offset positions in the original visible feature map according to the predicted offsets, thereby constructing a spatially more consistent aligned visible feature. It should be noted that the proposed method does not physically translate visible feature points. Instead, it regenerates the aligned visible feature map through backward sampling. Therefore, the predicted offsets are more suitable to be interpreted as task-oriented feature sampling offsets rather than a strict image-level geometric registration field. Although the above formulation omits the scale and subband indices, the same alignment procedure is applied at all selected feature scales. OffsetNet parameters are shared only across the four subbands within each scale and are independent across scales.

3.4 Swin Transformer-based Cross-modal Feature Fusion Module

After spatial alignment by OffsetNet, the aligned visible subband feature $F^{\text{rgb-align}}$ and the infrared subband feature F^{ir} are spatially more consistent. However, spatial alignment alone is insufficient to fully exploit the complementary information between the two modalities. Therefore, a Swin Transformer-based cross-modal feature fusion module, termed SwinFusion, is further introduced to model intra-modal contextual information and perform inter-modal feature fusion. SwinFusion first performs intra-modal feature modeling for the aligned visible subband feature and infrared subband

feature, and then performs inter-modal feature fusion. Within each feature scale, the same SwinFusion instance is applied to all four wavelet subbands, while different scales use independent instances.

The proposed SwinFusion adopts an Intra-Inter Swin Fusion design, which is motivated by previous multispectral detection studies that model intra-modality and inter-modality relationships for feature fusion [7, 9]. Specifically, the aligned visible and infrared features are first independently fed into Swin units for intra-modal feature modeling, and then cross-modal feature fusion is performed between the two modality-specific representations.

The intra-modal feature modeling process can be formulated as:

$$F_{\text{intra}}^{\text{rgb}} = \mathcal{S}_{\text{intra}}^{\text{rgb}}(F^{\text{rgb-align}}), \quad F_{\text{intra}}^{\text{ir}} = \mathcal{S}_{\text{intra}}^{\text{ir}}(F^{\text{ir}}), \quad (22)$$

where $\mathcal{S}_{\text{intra}}^{\text{rgb}}(\cdot)$ and $\mathcal{S}_{\text{intra}}^{\text{ir}}(\cdot)$ denote the intra-modal Swin units for visible and infrared features, respectively.

In the intra-modal stage, the query, key, and value are all derived from the same modality. Thus, visible self-attention aggregates contextual information within the aligned visible representation, while infrared self-attention independently models the infrared representation. This stage strengthens modality-specific features before information is exchanged between modalities.

In the second stage, inter-modal feature fusion is performed between the two modality-specific representations. The visible branch uses infrared features as complementary information, while the infrared branch uses visible features as complementary information:

$$Z^{\text{rgb}} = \mathcal{S}_{\text{inter}}^{\text{rgb}}(F_{\text{intra}}^{\text{rgb}}, F_{\text{intra}}^{\text{ir}}), \quad (23)$$

$$Z^{\text{ir}} = \mathcal{S}_{\text{inter}}^{\text{ir}}(F_{\text{intra}}^{\text{ir}}, F_{\text{intra}}^{\text{rgb}}), \quad (24)$$

where $\mathcal{S}_{\text{inter}}^{\text{rgb}}(\cdot)$ and $\mathcal{S}_{\text{inter}}^{\text{ir}}(\cdot)$ denote the inter-modal Swin fusion units.

Unlike intra-modal self-attention, inter-modal cross-attention derives the query from the modality being updated and the key and value from the complementary modality. Let X^{rgb} and X^{ir} denote the token representations of $F_{\text{intra}}^{\text{rgb}}$ and $F_{\text{intra}}^{\text{ir}}$ within a pair of spatially corresponding windows. For a single attention head, the projections are written as

$$\begin{aligned} Q^m &= X^m W_Q^m, \\ K^m &= X^m W_K^m, \\ V^m &= X^m W_V^m, \quad m \in \{\text{rgb}, \text{ir}\}, \end{aligned} \quad (25)$$

where W_Q^m , W_K^m , and W_V^m are learnable projection matrices. Head indices are omitted for clarity. The two directions of cross-modal attention are expressed as

$$A^{\text{rgb} \leftarrow \text{ir}} = \text{Softmax}\left(\frac{Q^{\text{rgb}}(K^{\text{ir}})^{\top}}{\sqrt{d}} + R^{\text{rgb} \leftarrow \text{ir}}\right) V^{\text{ir}}, \quad (26)$$

$$A^{\text{ir} \leftarrow \text{rgb}} = \text{Softmax}\left(\frac{Q^{\text{ir}}(K^{\text{rgb}})^{\top}}{\sqrt{d}} + R^{\text{ir} \leftarrow \text{rgb}}\right) V^{\text{rgb}}, \quad (27)$$

where d is the key dimension of one attention head, and R denotes the relative positional bias together with the attention mask when applicable. Softmax is applied over the key tokens. In the visible update, each visible query retrieves a weighted combination of infrared value tokens; in the infrared update, each infrared query retrieves complementary visible information. The outputs of multiple heads are concatenated and projected within the corresponding inter-modal Swin unit to produce the enhanced branch representations Z^{rgb} and Z^{ir} . The two directions therefore preserve their respective query coordinate systems while exchanging modality-dependent information.

According to the structure shown in Fig. 5, the enhanced features from the two modalities are concatenated and projected by a 1×1 convolution to generate the fused subband feature:

$$F^{\text{fuse}} = \text{Conv}_{1 \times 1} \left(\text{Concat} \left(Z^{\text{rgb}}, Z^{\text{ir}} \right) \right). \quad (28)$$

The window mechanism is used in both stages, but the sources of the query, key, and value differ: the intra-modal stage uses window-based self-attention, whereas the inter-modal stage uses window-based cross-attention between the two modalities. In the inter-modal stage, the visible and infrared features follow the same window partition so that each query window interacts with its spatially corresponding key and value window. For shifted-window attention, both modalities use the same spatial shift, and an attention mask prevents invalid interactions introduced by cyclic shifting. The shift is reversed when the window outputs are restored to the feature-map layout. Alternating regular and shifted windows enables information exchange across neighboring windows while avoiding attention over all spatial tokens at once.

This design separates the roles of alignment and fusion. OffsetNet first resamples visible features into the infrared reference coordinate space, whereas cross-attention subsequently aggregates complementary information according to learned feature affinities within the corresponding windows. The attention weights describe feature interactions rather than explicit geometric displacements. By first modeling intra-modal context and then performing bidirectional inter-modal interaction, SwinFusion provides a structured mechanism for combining the aligned visible and infrared representations.

Although the above formulation omits the scale and subband indices, a separate SwinFusion instance is used at each selected feature scale. Within that scale, its parameters are shared across the LL, LH, HL, and HH subbands to generate the corresponding fused subband features; parameters are not shared across scales.

In the ablation study, several Swin-based fusion variants are compared to verify the effectiveness of the proposed Intra-Inter Swin Fusion strategy. The compared variants include Intra Swin Fusion, Inter Swin Fusion, Inter-Intra Swin Fusion, and the proposed Intra-Inter Swin Fusion. Their detailed structures and quantitative comparisons are presented in the ablation study.

3.5 Wavelet-based Neck and Detection Head

After cross-modal alignment and feature fusion, the fused wavelet subbands are passed to the neck shown in Fig. 2 for multi-scale feature aggregation. The neck integrates IDWT reconstruction with feature aggregation and produces three detection features,

denoted by F_3^{neck} , F_4^{neck} , and F_5^{neck} . These features are then fed into the YOLO detection head to predict object categories, bounding boxes, and confidence scores:

$$Y = \mathcal{D}(\{F_3^{\text{neck}}, F_4^{\text{neck}}, F_5^{\text{neck}}\}). \quad (29)$$

4 Experiments

This section presents the experimental validation and result analysis of the proposed visible-infrared object detection network. We first introduce the datasets and implementation details used in the experiments. Then, the proposed method is compared with existing visible-infrared object detection methods to verify its overall detection performance. Furthermore, ablation studies are conducted to analyze the effectiveness of OffsetNet alignment network, and SwinFusion cross-modal feature enhancement module. Finally, detection results and offset visualizations are provided to further analyze the alignment and fusion effects of the proposed method.

4.1 Datasets

To comprehensively evaluate the effectiveness of the proposed method in different scenarios, experiments are conducted on four public visible-infrared object detection datasets, namely FLIR, LLVIP, VEDAI, and DroneVehicle. These datasets cover typical application scenarios such as autonomous driving, low-light pedestrian detection, aerial small-object detection, and UAV-based vehicle detection. In addition, the visible and infrared images in different datasets vary in imaging conditions, object scales, background complexity, and spatial alignment degree. Therefore, these datasets can comprehensively evaluate the effectiveness of the proposed method in cross-modal feature fusion, spatial alignment, and object detection under complex scenarios.

1) FLIR: The FLIR ADAS dataset is a commonly used visible-infrared object detection dataset for autonomous driving scenarios, containing paired visible and infrared images. The dataset covers complex scenes, including daytime, nighttime, and various road environments. Since the visible and infrared images in the original FLIR dataset are not strictly aligned, the manually aligned FLIR dataset is adopted in this work [23]. This dataset contains 4129 image pairs for training and 1013 image pairs for testing, including only three object categories: person, car, and bicycle.

2) LLVIP: The LLVIP [24] dataset is a visible-infrared pedestrian detection dataset designed for low-light scenarios, mainly collected at night and under weak illumination conditions. It contains 15488 visible-infrared image pairs, including 12025 pairs for training and 3463 pairs for testing, with pedestrian annotations. Under low-light conditions, visible images usually suffer from insufficient brightness, low contrast, and blurred details, while infrared images can provide relatively stable thermal radiation responses. Therefore, this dataset is suitable for evaluating the ability of the model to exploit infrared information for improving pedestrian detection performance in low-light environments.

3) VEDAI: The VEDAI [25] dataset is an aerial visible-infrared vehicle detection dataset, containing 1210 paired visible and infrared images, including 1089 pairs for training and 121 pairs for testing. Different from ground-view datasets, the objects in

VEDAI are usually small and are easily affected by complex backgrounds, viewpoint variations, and dense object distributions. The dataset contains eight vehicle-related categories, including car, truck, tractor, pickup, van, boat, camping car, and plane.

4) DroneVehicle: The DroneVehicle [26] dataset is a large-scale UAV-view visible-infrared object detection dataset, containing 28439 paired visible and infrared images collected by UAV platforms, including 19459 pairs for training and 8980 pairs for testing. This dataset has the characteristics of significant viewpoint variations, large object scale differences, complex backgrounds, and diverse illumination conditions. It contains five vehicle categories: car, truck, bus, van, and freight car. Due to UAV motion and the imaging differences between different sensors, certain spatial discrepancies may exist between visible and infrared images.

4.2 Implementation Details

Haar wavelets are used for feature decomposition and reconstruction. SwinFusion uses 4×4 attention windows with eight heads and an MLP expansion ratio of 2. Each modality-specific intra-modal stage and the bidirectional inter-modal stage contain two successive attention blocks, using regular windows followed by shifted windows with a shift size of 2. One Intra-Inter fusion group is used at each feature scale. Layer normalization and GELU activation are used in the Swin blocks, and dropout, attention dropout, and stochastic depth are set to zero. Module parameters are shared across the four subbands within each scale and are independent across scales.

The proposed method is implemented based on PyTorch and the YOLO detection framework, and all experiments are conducted on a single NVIDIA RTX 4090 GPU. All input images are resized to 640×640 , and the batch size is set to 8. The model is trained for 100 epochs using the SGD optimizer with an initial learning rate of 1×10^{-2} and a weight decay of 5×10^{-4} . In addition, warm-up and cosine annealing learning rate scheduling are adopted during training. Data augmentation strategies, including random scaling, random translation, random flipping, color jittering, and Mosaic augmentation, are used to improve the generalization ability of the model. Except for the settings explicitly specified above, all remaining training settings follow the default configuration of the YOLOv11 implementation used in this work. Unless otherwise specified, the same training configuration is adopted for experiments on different datasets. In ablation studies, only the module to be evaluated is changed, while the other training settings remain unchanged to ensure fair comparisons.

The original YOLOv11 detection loss and its default weighting are used without modification. OffsetNet and SwinFusion are optimized end-to-end through this detection objective. No additional alignment loss, offset regularization term, or separate supervision for the Auxiliary Branch is introduced.

Inference time is measured on the same NVIDIA RTX 4090 GPU using an inference batch size of 4. Apart from the settings explicitly specified in this work, inference uses the default configuration of the YOLOv11 implementation.

4.3 Comparison With Existing Methods

To verify the overall effectiveness of the proposed method, we compare the proposed visible-infrared object detection network with representative multimodal object detection methods in recent years. The compared methods include typical dual-stream fusion detection methods, attention-based cross-modal fusion methods, and visible-infrared detection methods based on Transformer or Mamba structures. Specifically, CMAF, CFT, ICAFusion, and WaveMamba are retrained and evaluated under the experimental settings used in this work and are marked with $\dagger$ in Tables 1–4. Results for the other comparison methods are taken from their corresponding publications. Comparisons with these published results provide context, but may also reflect differences in the original training and evaluation protocols.

Experiments are conducted on four public datasets, including FLIR, LLVIP, VEDAI, and DroneVehicle. The main evaluation metrics are mAP_{50} and mAP . mAP_{50} denotes the mean average precision at an IoU threshold of 0.5, while mAP denotes the average mAP under IoU thresholds from 0.5 to 0.95 with a step size of 0.05. The quantitative comparison results on each dataset are presented as follows.

Quantitative results on the FLIR dataset: Table 1 presents the quantitative comparison between the proposed method and existing methods on the FLIR dataset. It can be observed that the proposed method achieves competitive detection performance in terms of both mAP_{50} and mAP . Compared with simple feature concatenation or convolution-based fusion methods, the proposed method performs feature-level spatial alignment on visible features through OffsetNet before fusion, alleviating the negative influence of cross-modal spatial misalignment. Furthermore, the SwinFusion module models intra-modal contextual relationships and inter-modal complementary relationships, making the fused features more discriminative.

Table 1 Quantitative results for the FLIR dataset

Method	Year	mAP_{50} (%)	mAP (%)
GAFF [9]	2021	72.7	37.3
CFT [5] †	2022	75.2	39.7
CMAF [11] †	2022	75.8	41.1
UA-CMDet [26]	2022	78.6	-
CSSA [27]	2023	79.2	41.3
MFPT [28]	2023	80.0	-
ICAFusion [7] †	2024	76.5	41.2
CrossFormer [29]	2024	79.3	42.1
WaveMamba [17] †	2025	79.4	42.4
Ours	-	80.9	42.8

† denotes results reproduced in this work; unmarked comparison methods use results reported in the corresponding publications. “-” denotes an unavailable entry. Bold values indicate the best results in each metric column.

Quantitative results on the LLVIP dataset: Table 2 shows the experimental results on LLVIP. The proposed method achieves 97.5% mAP_{50} and 66.1% mAP , matching WaveMamba in mAP_{50} and exceeding it by 0.1 percentage points in mAP .

These results indicate comparable performance to WaveMamba on this low-light pedestrian detection dataset. The small numerical difference should not be interpreted as evidence of a statistically significant improvement without repeated-run evaluation.

Table 2 Quantitative results for the LLVIP dataset

Method	Year	mAP ₅₀ (%)	mAP (%)
GAFF [9]	2021	94.0	55.8
CFT [5] [†]	2022	96.9	65.1
UA-CMDet [26]	2022	96.3	-
CMAF [11] [†]	2022	97.1	65.8
CSSA [27]	2023	94.3	59.2
ICAFusion [7] [†]	2024	97.3	65.3
M ² FNet [30]	2024	96.9	-
CrossFormer [29]	2024	97.4	65.1
WaveMamba [17] [†]	2025	97.5	66.0
Ours	-	97.5	66.1

[†] denotes results reproduced in this work; unmarked comparison methods use results reported in the corresponding publications. “-” denotes an unavailable entry. Bold values indicate the best results in each metric column.

Quantitative results on the VEDAI dataset:

Table 3 presents the comparison results of different methods on the VEDAI dataset. VEDAI is an aerial visible-infrared dataset in which objects are usually small and easily affected by complex backgrounds and viewpoint variations. As shown in Table 3, the proposed method achieves the best mAP₅₀ of 78.2%, outperforming WaveMamba by 0.3 percentage points. This indicates that the proposed spatial alignment and cross-modal fusion strategy is effective for detecting aerial small objects under the IoU threshold of 0.5.

However, the mAP of the proposed method is 49.8%, which is lower than that of WaveMamba. This indicates better detection performance at an IoU threshold of 0.5, but lower performance when averaged over IoU thresholds from 0.5 to 0.95. The difference alone does not identify separate changes in recall or confidence, and suggests room for improving localization at stricter IoU thresholds. One possible reason is that small objects in VEDAI are more sensitive to slight localization deviations, and feature-level alignment may introduce local sampling variations when the original visible-infrared image pairs are already relatively well aligned. Nevertheless, the results demonstrate that the proposed method has competitive performance on aerial visible-infrared object detection, especially in terms of mAP₅₀.

Quantitative results on the DroneVehicle dataset: Table 4 shows the quantitative comparison results of different methods on the DroneVehicle dataset. The DroneVehicle dataset has significant object scale variations, complex viewpoints, strong background interference, and cross-modal spatial discrepancies, providing an additional evaluation of the proposed method in UAV-based multimodal object detection scenarios. It can be seen that the proposed method still achieves favorable

Table 3 Quantitative comparison of different methods on the VEDAI dataset

Method	Year	mAP ₅₀ (%)	mAP (%)
CFT [5] [†]	2022	71.7	43.3
CMAF [11] [†]	2022	71.6	46.4
SuperYOLO [3]	2023	73.6	-
ICAFusion [7] [†]	2024	76.6	44.9
WaveMamba [17] [†]	2025	77.9	51.3
Ours	-	78.2	49.8

[†] denotes results reproduced in this work; unmarked comparison methods use results reported in the corresponding publications. “-” denotes an unavailable entry. Bold values indicate the best results in each metric column.

detection performance under complex UAV-view scenarios. Compared with conventional dual-stream fusion methods, the proposed method adaptively resamples visible features through OffsetNet, making them better match the infrared feature space at the feature level. Meanwhile, SwinFusion further enhances the complementary information between the two modalities through intra-modal fusion and inter-modal interaction. Therefore, the proposed method can maintain strong detection capability in scenarios with large object scale variations and complex backgrounds.

Table 4 Quantitative comparison of different methods on the DroneVehicle dataset

Method	Year	mAP ₅₀ (%)	mAP (%)
CFT [5] [†]	2022	82.5	63.0
CMAF [11] [†]	2022	82.0	62.5
ICAFusion [7] [†]	2024	82.4	63.1
DEYOLO [31]	2024	81.8	61.4
WaveMamba [17] [†]	2025	82.8	63.5
Ours	-	83.4	64.0

[†] denotes results reproduced in this work; unmarked comparison methods use results reported in the corresponding publications. “-” denotes an unavailable entry. Bold values indicate the best results in each metric column.

Figure 7 compares representative detections from CFT, CMAF, ICAFusion, and the proposed method on FLIR, DroneVehicle, LLVIP, and VEDAI. In the selected examples, the proposed method detects objects missed by some comparison methods and produces bounding boxes that more closely match the ground truth in several regions. These examples illustrate the detection behavior of the proposed framework across the evaluated scenarios; they do not constitute a cross-dataset generalization experiment.

Overall, the experimental results on the four datasets show that the proposed method achieves favorable detection performance in different scenarios. For datasets with cross-modal spatial discrepancies, OffsetNet can alleviate the inconsistency

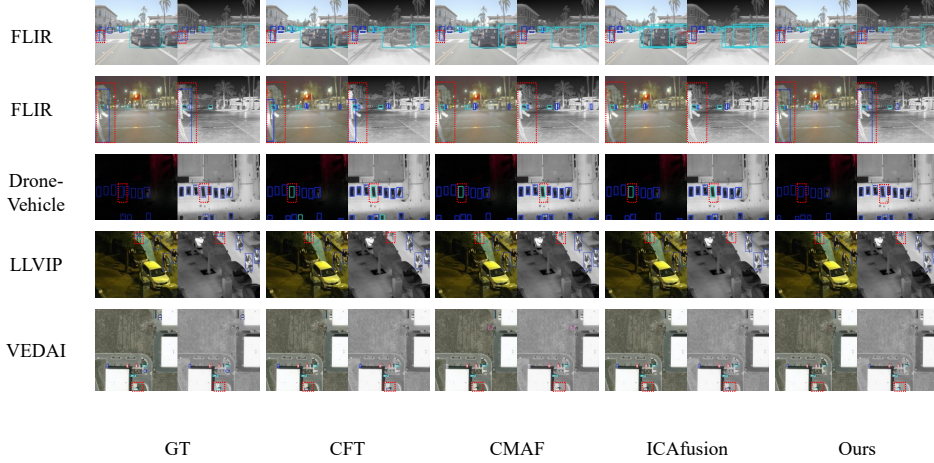


Fig. 7 Qualitative detection results on multiple visible-infrared object detection datasets. The columns correspond to ground truth, CFT, CMAF, ICAfusion, and the proposed method, respectively. The rows show representative samples from FLIR, DroneVehicle, LLVIP, and VEDAI.

between visible and infrared features through feature-level resampling. For small-object and complex-background scenarios, the adopted wavelet-based feature modeling operation helps preserve object edges and local details. For scenarios with strong modality complementarity, SwinFusion further enhances intra-modal contextual representation and inter-modal information interaction. These results demonstrate that the proposed network can effectively improve visible-infrared object detection performance.

4.4 Ablation Studies

To further verify the effectiveness of each component, overall component and feature-fusion ablations are conducted on FLIR, while OffsetNet ablations are conducted on both FLIR and VEDAI. The baseline is defined as a dual-stream YOLO detector with conventional cross-modal feature fusion, without the adopted wavelet-based feature modeling operation, OffsetNet, and SwinFusion. On this basis, the adopted WaveMamba-inspired wavelet-based feature modeling operation, OffsetNet, and SwinFusion are gradually introduced to analyze their influence on detection accuracy and computational complexity. It should be noted that the backbone configuration is kept fixed for all ablation variants, and only the evaluated component is changed. Unless otherwise specified, all ablation experiments adopt the same dataset split, input size, training epochs, and optimization strategy to ensure fair comparisons.

4.4.1 Overall Component Ablation

To verify the contribution of each component in the proposed framework, a step-by-step ablation study is conducted on the FLIR dataset. It should be noted that the wavelet-based feature modeling operation used in this work is inspired by WaveMamba [17] and is not regarded as the main methodological novelty of this paper.

Nevertheless, since DWT and IDWT are integrated into the overall detection framework, this operation is included in the overall ablation study to analyze its influence on detection performance. The main task-specific modules proposed in this work are OffsetNet and SwinFusion. In addition to detection accuracy, the number of parameters and inference time are also reported to analyze the accuracy-complexity trade-off of different model variants. The backbone configuration is kept unchanged for all variants, so the comparison mainly reflects the influence of the adopted wavelet operation, OffsetNet, and SwinFusion.

As shown in Table 5, the baseline model achieves 74.3% mAP₅₀ and 39.1% mAP, with 54.48M parameters and 35.0 ms inference time. After introducing the adopted wavelet-based feature modeling operation, the detection performance improves to 76.7% mAP₅₀ and 40.5% mAP. This indicates that wavelet-based decomposition and reconstruction are beneficial for preserving structural and detail information during feature extraction and aggregation. Meanwhile, the inference time increases from 35.0 ms to 44.3 ms, showing that the adopted wavelet operation introduces additional computational overhead.

When OffsetNet is further introduced, the model achieves 78.0% mAP₅₀ and 40.9% mAP, demonstrating that feature-level spatial alignment can alleviate local misalignment between visible and infrared features. Compared with the wavelet-based variant, OffsetNet increases the parameter number from 50.01M to 53.24M and the inference time from 44.3 ms to 71.0 ms. The additional learnable parameters arise from the convolutional and projection layers, together with other trainable components introduced in the alignment module. Local correlation computation and bilinear feature sampling do not themselves introduce learnable parameters, but they contribute to the computational overhead alongside the learnable layers and thus affect inference time.

When SwinFusion is added without OffsetNet, the model achieves 78.6% mAP₅₀ and 41.6% mAP, indicating that Swin Transformer-based intra-modal modeling and inter-modal interaction are effective for enhancing cross-modal feature representations. However, SwinFusion also brings a larger computational cost, increasing the parameter number to 116.17M and the inference time to 112.8 ms.

Finally, when the adopted wavelet-based feature modeling operation, OffsetNet, and SwinFusion are jointly used, the model obtains the best detection performance, with 80.9% mAP₅₀ and 42.8% mAP. The complete model contains 119.40M parameters and requires 158.3 ms inference time. Although the full model introduces additional computational overhead, it achieves the highest detection accuracy among all variants. These results demonstrate that the adopted wavelet operation provides useful frequency-domain representations, while the proposed OffsetNet and SwinFusion further improve feature-level alignment and cross-modal feature fusion. Overall, the complete framework achieves the highest detection accuracy among the evaluated variants at the cost of increased parameter count and inference time.

To further provide an intuitive comparison, qualitative visualization results of different component variants are shown in Fig. 8. The baseline variant in Fig. 8(b) fails to detect the overlapping pedestrians. After introducing OffsetNet, the variant in Fig. 8(c) successfully detects one of the overlapping pedestrians, and its predicted

bounding box is more consistent with the object contour among the compared variants. This indicates that the feature-level spatial alignment provided by OffsetNet helps improve localization under local cross-modal misalignment. Although the result in Fig. 8(d) is less accurate than that in Fig. 8(c), the model with SwinFusion can still detect one overlapping pedestrian, benefiting from enhanced cross-modal feature interaction. Finally, the complete model in Fig. 8(e), which jointly employs OffsetNet and SwinFusion, produces a more complete and stable detection result among these variants.

Table 5 Overall component ablation results on the FLIR dataset.

Baseline	Adopted Wavelet	OffsetNet	SwinFusion	mAP ₅₀ (%)	mAP (%)	Params (M)	Inference Time (ms)
✓				74.3	39.1	54.48	35.0
✓	✓			76.7	40.5	50.01	44.3
✓	✓	✓		78.0	40.9	53.24	71.0
✓	✓		✓	78.6	41.6	116.17	112.8
✓	✓	✓	✓	80.9	42.8	119.40	158.3

“Adopted Wavelet” denotes the adopted WaveMamba-inspired wavelet-based feature modeling operation rather than a newly proposed module. Params and inference time are reported to analyze the accuracy–complexity trade-off of different model variants. Inference time is measured with batch size 4 on an NVIDIA RTX 4090; other unspecified inference settings follow the YOLOv11 defaults.

4.4.2 Ablation on the OffsetNet Alignment Network

To verify the role of OffsetNet in cross-modal feature alignment, different alignment strategies are compared on the FLIR and VEDAI datasets. These two datasets are selected because they present different degrees of cross-modal spatial misalignment. FLIR contains visible-infrared image pairs with relatively obvious local discrepancies, while VEDAI is relatively well aligned. Therefore, experiments on these two datasets can evaluate the effectiveness and robustness of OffsetNet under different alignment conditions.

Specifically, four settings are considered: without the alignment module, OffsetNet without the auxiliary branch, OffsetNet without the Main Branch, and the complete OffsetNet for feature-level spatial alignment. The experimental results are shown in Table 6.

Table 6 shows that retaining only the Main Branch improves FLIR performance relative to the variant without alignment, but reduces VEDAI performance from 74.3% to 73.5% mAP₅₀ and from 46.8% to 45.6% mAP. This suggests that offset prediction without local correlation guidance may introduce unnecessary feature perturbations when the image pairs are relatively well aligned. The Auxiliary Branch supplies matching cues rather than imposing an explicit constraint on the range or direction of the predicted offsets.

When only the Auxiliary Branch is retained, the model improves over the variant without alignment on both FLIR and VEDAI. On FLIR, it performs slightly below

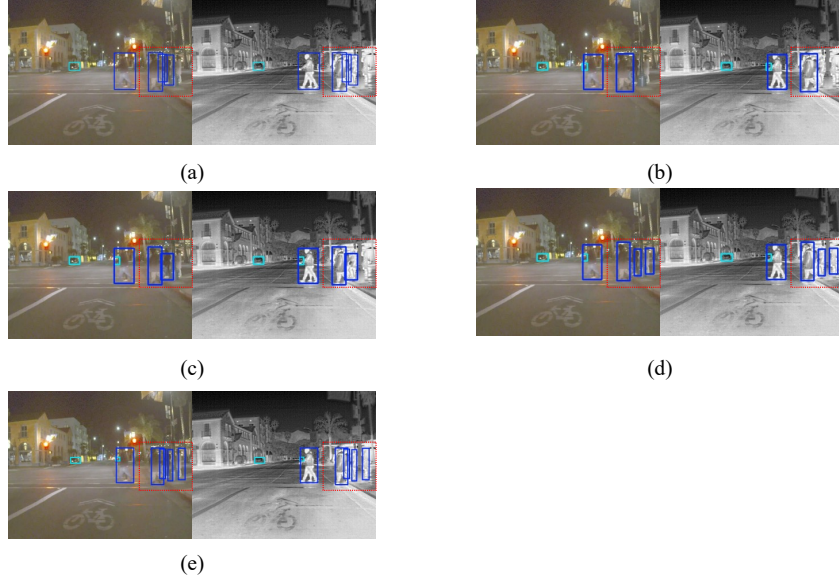


Fig. 8 Qualitative visualization of the overall component ablation on the FLIR dataset. (a) Ground truth. (b) Baseline with the adopted wavelet-based feature modeling. (c) Baseline with the adopted wavelet-based feature modeling and OffsetNet. (d) Baseline with the adopted wavelet-based feature modeling and SwinFusion. (e) Baseline with the adopted wavelet-based feature modeling, OffsetNet, and SwinFusion.

Table 6 Ablation results of the OffsetNet alignment network on the FLIR and VEDAI datasets.

Method	Main	Aux.	Align.	FLIR		VEDAI	
				mAP ₅₀ (%)	mAP (%)	mAP ₅₀ (%)	mAP (%)
w/o Alignment				78.6	41.6	74.3	46.8
w/o Auxiliary Branch	✓		✓	79.6	42.0	73.5	45.6
w/o Main Branch		✓	✓	79.2	41.8	75.1	48.7
OffsetNet	✓	✓	✓	80.9	42.8	78.2	49.8

“Main”, “Aux.”, and “Align.” denote the main branch, auxiliary branch, and offset-guided feature alignment, respectively. Bold values indicate the best results.

the Main-Branch-only variant, whereas on VEDAI it performs better. These results suggest that the two branches provide different cues for offset prediction. Neither single-branch variant matches the complete OffsetNet on these datasets.

To further analyze the behavior of OffsetNet, offset visualization results are shown in Fig. 9. At each spatial position of the P3 subband feature grid, the predicted offsets are averaged arithmetically across the four subbands before color mapping:

$$\bar{d}_x(p) = \frac{1}{4} \sum_{s \in \mathcal{S}} d_x^{(s)}(p), \quad \bar{d}_y(p) = \frac{1}{4} \sum_{s \in \mathcal{S}} d_y^{(s)}(p), \quad \mathcal{S} = \{LL, LH, HL, HH\}. \quad (30)$$

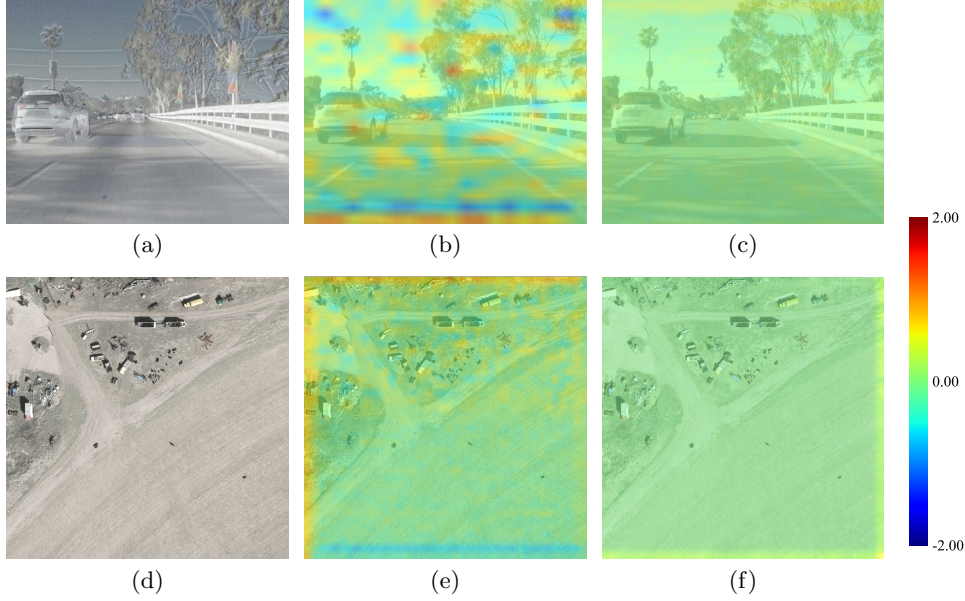


Fig. 9 Offset visualization at the P3 scale, using the arithmetic mean across the four wavelet subbands. (a)–(c): FLIR; (d)–(f): VEDAI. (a), (d): RGB-IR overlays; (b), (e): mean horizontal offsets $\bar{d}_x$; (c), (f): mean vertical offsets $\bar{d}_y$. The shared color bar gives signed offsets in feature-grid units, with the same numerical mapping for both datasets.

The two heatmaps show the signed mean horizontal and vertical offsets, respectively. Values are expressed in feature-grid units, without clipping, and the same numerical color mapping is used for FLIR and VEDAI. To overlay the maps on the original images, only the heatmaps are spatially interpolated; offset values are not rescaled into original-image pixel units. These plots summarize the mean subband responses rather than the offset field of an individual subband. Oppositely signed offsets can cancel during averaging, so a weak mean response does not imply small offsets in every subband. Accordingly, the color values should not be interpreted as the magnitude of the two-dimensional offset vector. Instead, the sign indicates the sampling direction along the corresponding axis, and the absolute value reflects the displacement strength of that component. The spatial variation of d_x and d_y shows that OffsetNet learns location-dependent, task-oriented feature sampling offsets rather than applying a uniform translation or performing strict image-level geometric registration.

More specifically, the RGB-IR overlay of the selected FLIR sample suggests that the visible cross-modal discrepancy is predominantly horizontal. The d_x map exhibits a stronger response than the d_y map, which is qualitatively consistent with this observation. For the selected VEDAI sample, the image pair appears relatively well aligned, and the combined offset responses are comparatively weak, especially in d_y . These sample-level observations describe the learned sampling behavior, but do not establish the geometric accuracy of the offsets.

Noticeable offset responses also occur in some background regions. Such responses do not by themselves demonstrate an alignment error, because background structures may also exhibit genuine cross-modal displacement. In addition, the offsets are learned through the detection objective without explicit pixel-wise correspondence or optical-flow supervision, so background context useful for detection may influence the sampling locations. Modality-specific appearance differences, repetitive textures, edges, and shadows may introduce ambiguous matching cues, while convolutional receptive-field aggregation and heatmap interpolation may spread localized responses into neighboring regions. These are possible explanations rather than mechanisms established by the visualization alone. Pixel-level registration accuracy cannot be determined from these heatmaps and would require correspondence or displacement ground truth. Accordingly, the maps should be interpreted as detection-oriented feature-sampling responses.

Finally, the Main Branch and the Auxiliary Branch are jointly designed to form the complete OffsetNet. The Main Branch is used to extract local response discrepancies and complementary information from bimodal features, while the Auxiliary Branch provides local correlation guidance to improve the stability of offset prediction. Their combination achieves the best detection results among the alignment variants on both datasets, supporting the usefulness of combining gated feature weighting with local correlation guidance. Nevertheless, OffsetNet is designed for task-oriented feature-level alignment rather than precise pixel-level registration. Therefore, the predicted offsets should be interpreted as adaptive feature sampling offsets for improving object detection, instead of accurate geometric displacement fields. The experimental results verify the effectiveness of the complete OffsetNet in feature-level spatial alignment.

4.4.3 Ablation on Different Feature Fusion Methods

To verify the effectiveness of the proposed SwinFusion module, different cross-modal feature fusion methods are compared on the FLIR dataset. Five fusion strategies are considered in this ablation study, including Concat, Intra Swin Fusion, Inter Swin Fusion, Inter-Intra Swin Fusion, and the proposed Intra-Inter Swin Fusion.

Concat directly stacks visible and infrared features along the channel dimension and serves as the basic fusion baseline. Intra Swin Fusion independently feeds visible and infrared features into Swin units for modality-specific contextual modeling, and then concatenates the two enhanced features followed by a 1×1 convolution to generate the fused feature. Inter Swin Fusion directly performs cross-modal feature interaction between visible and infrared features. Inter-Intra Swin Fusion first conducts inter-modal feature interaction and then performs intra-modal contextual modeling. In contrast, the proposed Intra-Inter Swin Fusion first enhances modality-specific contextual representations and then performs inter-modal feature fusion.

Figure 10 illustrates these four Swin-based variants, and Table 7 reports their detection performance together with the Concat baseline.

As shown in Table 7, the proposed Intra-Inter Swin Fusion achieves the best detection performance among all compared fusion strategies, with 80.9% mAP₅₀ and 42.8% mAP. Compared with the simple Concat baseline, Intra-Inter Swin Fusion improves mAP₅₀ from 78.0% to 80.9% and mAP from 40.9% to 42.8%, demonstrating the

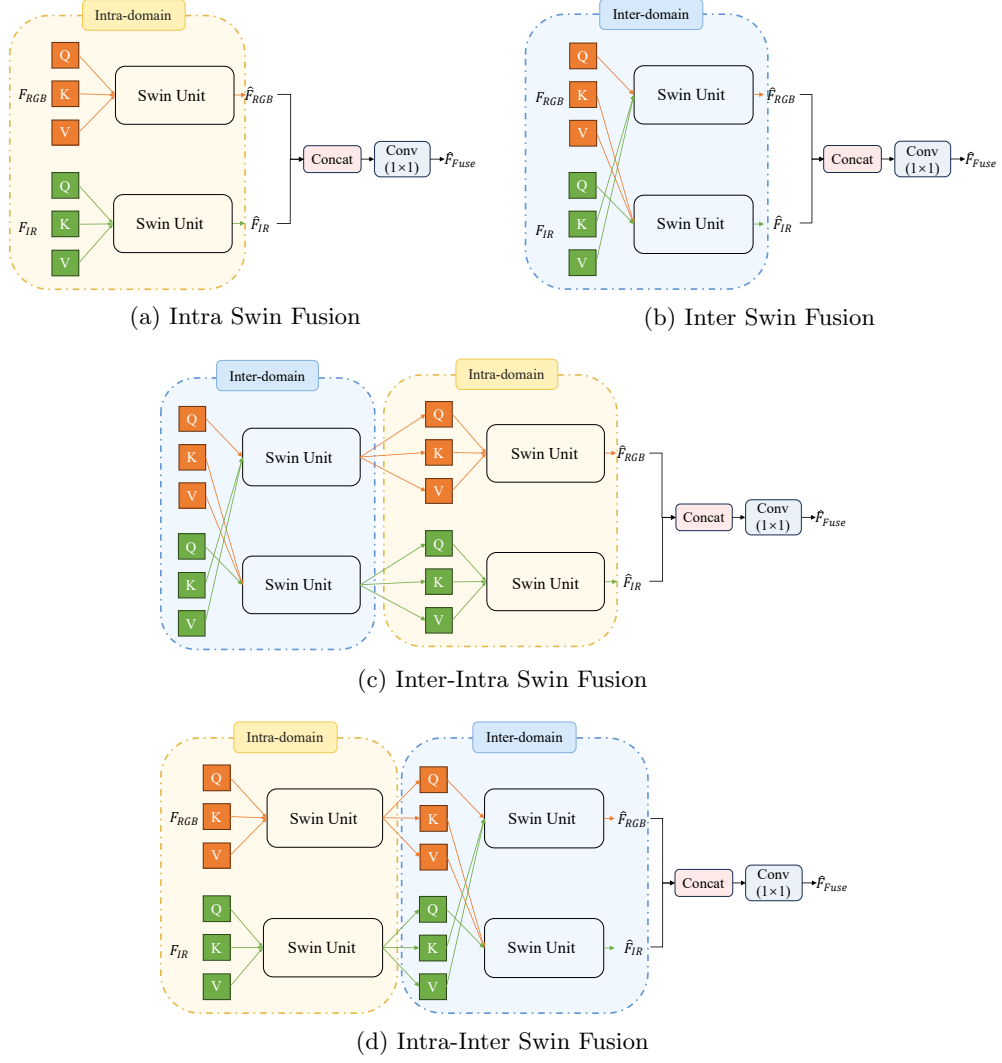


Fig. 10 Swin-based fusion variants: (a) Intra, (b) Inter, (c) Inter-Intra, and (d) Intra-Inter. Variant (d) is used in the proposed SwinFusion module.

Table 7 Ablation results of different Swin-based feature fusion methods on the FLIR dataset

Fusion Method	mAP ₅₀ (%)	mAP (%)
Concat	78.0	40.9
Intra Swin Fusion	79.8	42.3
Inter Swin Fusion	79.3	41.6
Inter-Intra Swin Fusion	78.7	41.3
Intra-Inter Swin Fusion	80.9	42.8

effectiveness of Swin Transformer-based feature interaction for visible-infrared feature fusion.

Compared with Intra Swin Fusion, the proposed Intra-Inter Swin Fusion further improves mAP_{50} from 79.8% to 80.9% and mAP from 42.3% to 42.8%. This indicates that intra-modal contextual modeling alone is insufficient to fully exploit the complementary information between visible and infrared modalities. Compared with Inter Swin Fusion, which directly performs cross-modal interaction, the proposed method also achieves better performance, suggesting that directly conducting inter-modal fusion without sufficient intra-modal feature enhancement may lead to less discriminative cross-modal representations.

In addition, Inter-Intra Swin Fusion obtains 78.7% mAP_{50} and 41.3% mAP. Reversing this order to Intra-Inter improves the two metrics by 2.2 and 1.5 percentage points, respectively. This comparison supports performing intra-modal modeling before inter-modal interaction in the evaluated framework. A possible explanation is that modality-specific contextual modeling provides more informative features for subsequent cross-attention. The observed accuracy difference supports the selected ordering but does not directly measure feature stability.

5 Conclusion

In this paper, we propose a dual-stream visible-infrared object detection network with spatial alignment and Swin Transformer-based cross-modal feature fusion. The proposed method extracts multi-scale features from visible and infrared images through a dual-stream backbone. Inspired by WaveMamba, wavelet-based feature modeling is adopted as an auxiliary frequency-domain operation in the backbone and neck to preserve structural and detail information during feature extraction and aggregation.

To alleviate feature-level spatial misalignment between visible and infrared modalities, OffsetNet is introduced to predict task-oriented sampling offsets and align visible features to the infrared reference feature space. Furthermore, SwinFusion is designed to perform intra-modal feature modeling and inter-modal feature fusion, enabling effective exploitation of complementary information between the two modalities. Experimental results on multiple visible-infrared object detection datasets demonstrate that the proposed method achieves competitive detection performance compared with existing methods. Ablation studies further verify the effectiveness of OffsetNet and SwinFusion, while visualization results show that the proposed method can produce more accurate detections and meaningful feature-level offset responses.

Although the proposed method achieves competitive performance, several limitations remain. First, OffsetNet focuses on task-oriented feature-level alignment rather than explicit global geometric registration, and its robustness under severe cross-modal displacement still needs further improvement. Second, the complete model has a relatively large number of parameters due to the dual-stream backbone and Swin Transformer-based fusion structure. Although SwinFusion improves cross-modal feature interaction, it inevitably introduces additional parameters and computational cost compared with simple convolutional fusion modules. This may limit the deployment of the proposed method in resource-constrained real-time scenarios. Third, the

experiments are mainly conducted on public visible-infrared datasets, and further validation under more diverse sensor settings and real-world deployment scenarios is still required. Nevertheless, the proposed framework demonstrates an effective alignment-before-fusion paradigm and achieves superior or competitive performance compared with existing visible-infrared object detection methods.

In future work, we will further improve the robustness of the alignment module under severe cross-modal misalignment and investigate more lightweight fusion designs, model compression strategies, and efficient backbone configurations to reduce the parameter size and computational cost for real-time deployment.

Acknowledgements. The authors would like to thank the providers of the public visible-infrared datasets used in this study.

Statements and Declarations

Competing Interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author contributions. Jingtao Hu: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Visualization, Writing – original draft. Yunxiang Zhu: Validation, Data curation, Writing – review & editing. Yangkun Zhu: Validation, Data curation, Writing – review & editing. Zhenzhou Shao: Writing – review & editing. Donghan Kim: Writing – review & editing. Junying Yang: Supervision, Project administration, Funding acquisition, Writing – review & editing. Yibo Zou: Conceptualization, Supervision, Project administration, Funding acquisition, Writing – review & editing.

Data availability. The datasets used in this study are publicly available. The FLIR, LLVIP, VEDAI, and DroneVehicle datasets can be accessed from their official websites or corresponding publications. The processed experimental results are available from the corresponding author upon reasonable request.

References

- [1] Jocher, G., et al.: ultralytics/yolov5: v7.0 – YOLOv5 SOTA Realtime Instance Segmentation. Zenodo (2022). <https://doi.org/10.5281/zenodo.7347926>
- [2] Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLOv8. <https://github.com/ultralytics/ultralytics> (2023)
- [3] Zhang, J., Lei, J., Xie, W., Fang, Z., Li, Y., Du, Q.: SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing* **61**, 1–15 (2023) <https://doi.org/10.1109/TGRS.2023.3258666>

- [4] Zhou, K., Chen, L., Cao, X.: Improving multispectral pedestrian detection by addressing modality imbalance problems. In: *Computer Vision – ECCV 2020*, pp. 787–803. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58523-5_46
- [5] Fang, Q., Han, D., Wang, Z.: Cross-modality fusion transformer for multispectral object detection. *SSRN Electronic Journal* (2022) <https://doi.org/10.2139/ssrn.4227745>
- [6] Zhang, L., Zhu, X., Chen, X., Yang, X., Lei, Z., Liu, Z.: Weakly aligned cross-modal learning for multispectral pedestrian detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5126–5136 (2019). <https://doi.org/10.1109/ICCV.2019.00523>
- [7] Shen, J., Chen, Y., Liu, Y., Zuo, X., Fan, H., Yang, W.: ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition* **145**, 109913 (2024) <https://doi.org/10.1016/j.patcog.2023.109913>
- [8] Liu, J., Zhang, S., Wang, S., Metaxas, D.N.: Multispectral deep neural networks for pedestrian detection. *arXiv preprint arXiv:1611.02644* (2016)
- [9] Zhang, H., Fromont, E., Lefèvre, S., Avignon, B.: Guided attentive feature fusion for multispectral pedestrian detection. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 72–80 (2021)
- [10] Dasgupta, K., Das, A., Das, S., Bhattacharya, U., Yogamani, S.: Spatio-contextual deep network-based multimodal pedestrian detection for autonomous driving. *IEEE transactions on intelligent transportation systems* **23**(9), 15940–15950 (2022)
- [11] Fang, Q., Wang, Z.: Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. *Pattern Recognition* **130**, 108786 (2022)
- [12] Li, Q., Zhang, C., Hu, Q., Fu, H., Zhu, P.: Confidence-aware fusion using Dempster-Shafer theory for multispectral pedestrian detection. *IEEE Transactions on Multimedia* **25**, 3420–3431 (2023)
- [13] Chen, C., Qi, J., Liu, X., Bin, K., Fu, R., Hu, X., Zhong, P.: Weakly misalignment-free adaptive feature alignment for UAVs-based multimodal object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26836–26845 (2024)
- [14] Qian, J., Qiao, B., Zhang, Y., Liu, T., Wang, S., Wu, G., Han, D.: DACFusion: Dual asymmetric cross-attention guided feature fusion for multispectral object detection. *Neurocomputing* **635**, 129913 (2025)
- [15] Hu, J., Ni, L., Peng, B., Li, T.: WtCAFNet: A wavelet transform and cross-attention modality-adaptive fusion network for multispectral object detection.

- [16] Wang, X., Chai, L., Chen, J., Shi, Z.: WCCNet: Wavelet-context cooperative network for efficient multispectral pedestrian detection. arXiv preprint arXiv:2308.01042 (2023)
- [17] Zhu, H., Dong, W., Yang, L., Li, H., Yang, Y., Ren, Y., Zhu, Q., Feng, Z., Li, C., Lin, S., *et al.*: WaveMamba: Wavelet-driven Mamba fusion for RGB-infrared object detection. In: 2025 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 11219–11229 (2025). IEEE
- [18] Li, C., Peng, X.: Joint Transformer and Mamba fusion for multispectral object detection. Image and Vision Computing **156**, 105468 (2025)
- [19] Hu, K., He, Y., Li, Y., Zhao, J., Chen, S., Kang, Y.: EI²Det: Edge-guided illumination-aware interactive learning for visible-infrared object detection. IEEE Transactions on Circuits and Systems for Video Technology **35**(7), 7101–7115 (2025) <https://doi.org/10.1109/TCSVT.2025.3539625>
- [20] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9992–10002 (2021). IEEE
- [21] Shi, D.: TransNeXt: Robust foveal visual perception for vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 17773–17783 (2024)
- [22] Sun, D., Yang, X., Liu, M.-Y., Kautz, J.: PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8934–8943 (2018). <https://doi.org/10.1109/CVPR.2018.00931>
- [23] Zhang, H., Fromont, E., Lefèvre, S., Avignon, B.: Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In: Proceedings of the IEEE International Conference on Image Processing, pp. 276–280 (2020). <https://doi.org/10.1109/ICIP40778.2020.9191080>
- [24] Jia, X., Zhu, C., Li, M., Tang, W., Zhou, W.: LLVIP: A visible-infrared paired dataset for low-light vision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3496–3504 (2021)
- [25] Razakarivony, S., Jurie, F.: Vehicle detection in aerial imagery: A small target detection benchmark. Journal of Visual Communication and Image Representation **34**, 187–203 (2016) <https://doi.org/10.1016/j.jvcir.2015.11.002>
- [26] Sun, Y., Cao, B., Zhu, P., Hu, Q.: Drone-based RGB-infrared cross-modality

- vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology* **32**(10), 6700–6713 (2022) <https://doi.org/10.1109/TCSVT.2022.3168279>
- [27] Cao, Y., Bin, J., Hamari, J., Blasch, E., Liu, Z.: Multimodal object detection by channel switching and spatial attention. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 403–411 (2023)
- [28] Zhu, Y., Sun, X., Wang, M., Huang, H.: Multi-modal feature pyramid transformer for RGB-infrared object detection. *IEEE Transactions on Intelligent Transportation Systems* **24**(9), 9984–9995 (2023) <https://doi.org/10.1109/TITS.2023.3266487>
- [29] Lee, S., Park, J., Park, J.: CrossFormer: Cross-guided attention for multi-modal object detection. *Pattern Recognition Letters* **179**, 144–150 (2024) <https://doi.org/10.1016/j.patrec.2024.02.012>
- [30] Jiang, C., Ren, H., Yang, H., Huo, H., Zhu, P., Yao, Z., Li, J., Sun, M., Yang, S.: M²FNet: Multi-modal fusion network for object detection from visible and thermal infrared images. *International Journal of Applied Earth Observation and Geoinformation* **130**, 103918 (2024) <https://doi.org/10.1016/j.jag.2024.103918>
- [31] Chen, Y., Wang, B., Guo, X., Zhu, W., He, J., Liu, X., Yuan, J.: DEYOLO: Dual-feature-enhancement YOLO for cross-modality object detection. In: *International Conference on Pattern Recognition*, pp. 236–252 (2024). Springer

Author Biographies



Jingtao Hu received the B.S. degree from Soochow University, Suzhou, China, in 2024. He is currently pursuing the M.S. degree with the School of Optoelectronic Science and Technology, Soochow University. His research interests include deep learning, machine vision, multimodal object detection, visible-infrared feature alignment, and cross-modal feature fusion.



Yunxiang Zhu received the B.S. degree in electronic information science and technology from Soochow University, Suzhou, China, in 2023. He is currently pursuing the M.S. degree with the School of Optoelectronic Science and Technology, Soochow University. His research interests include computer vision, image processing, deep learning, and multimodal object detection.



Yangkun Zhu received the B.S. degree from Soochow University, Suzhou, China, in 2024. He is currently pursuing the M.S. degree with the School of Optoelectronic Science and Technology, Soochow University. His research interests include deep learning, machine vision, image processing, multimodal object detection, and cross-modal feature fusion.



Zhenzhou Shao received the B.E. and M.E. degrees in information engineering from Northeastern University, China, in 2007 and 2009, respectively, and the Ph.D. degree in mechanical engineering from the University of Tennessee, Knoxville, USA, in 2013. He is currently a Professor with the College of Information Engineering, Capital Normal University, China. His research interests include intelligent robotics, control systems, image processing, and computer vision.



Donghan Kim is a Professor with the Department of Electronic Engineering at Kyung Hee University, Republic of Korea. He received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology. He was a postdoctoral fellow at the University of Illinois Urbana-Champaign and previously worked as a Senior Engineer at Samsung Electronics. His research interests include human-robot interaction, service robotics, cooperative robotics, multi-robot systems, and robotic perception.



Junying Yang received the Ph.D. degree from the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China, in 2022. She is currently a Lecturer with the School of Optoelectronic Science and Technology, Soochow University, Suzhou, China. Her primary research interests include intelligent remote sensing signal processing, image processing, and their related applications.



Yibo Zou received the B.S. degree in power mechanical engineering from Jiangsu University, China, in 2010, and the M.S. and Ph.D. degrees in mechanical engineering from Leibniz University Hannover, Germany, in 2013 and 2017, respectively. He is currently an Associate Professor with the School of Optoelectronic Science and Technology, Soochow University, China. His research interests include optical metrology, image processing, computer vision, non-contact precision measurement, and robotic vision.