Electronic Research Archive

https://www.aimspress.com/journal/era

ERA, 34(2): 777–812.
DOI: 10.3934/era.2026035
Received: 22 October 2025
Revised: 05 January 2026
Accepted: 07 January 2026
Published: 23 January 2026

Research article

Tiny bird detection and location guided by heterogeneous binocular images in transformer substation scene

Qiyun Yin[1], Xiao Li[2], Chang Xu[3], Yanjie An[1] and Qingwu Li[3,*]

[1] Ultrahigh Voltage Company, State Grid Ningxia Electric Power Co., Ltd., Ningxia 750000, China

[2] College of Artificial Intelligence and Automation, Hohai University, Jiangsu 213000, China

[3] College of Information Science and Engineering, Hohai University, Jiangsu 213000, China

* **Correspondence:** Email: li_qingwu@163.com; Tel: +8613092520035.

Abstract: Bird activities like nesting and perching in transformer substations threaten power grid stability by causing short circuits and insulation failures. Existing bird repelling devices are inefficient due to the lack of accurate detection and positioning, leading to energy waste and safety hazards from continuous operation. To address this, this paper develops a tiny bird detection and location system guided by heterogeneous binocular images for precise, targeted repulsion. For well-lit scenarios, a two-stage contextual information enhancement network is proposed. It mines multiscale context to highlight tiny bird regions, fuses context with second-stage features via channel dimension enhancement, and uses spatial attention for accurate localization. For low-light or occluded scenes, a multiscale contextual feature enhancement network processes infrared images, adopting multibranch cross-level feature fusion and combining transformer with multisize convolution to suppress background and thermal radiation interference. Additionally, heterogeneous binocular cameras are calibrated to calculate bird spatial distance, integrating detection results with spatial information to drive a laser repelling device. Experimental results in real substation environments show the system meets engineering requirements for robustness and accuracy. The detection in visible images achieves an overall average precision of 59.8%, while the infrared detection outperforms advanced algorithms in key metrics. The spatial localization error is controlled within 4.9%, significantly improving bird expulsion success rate and reducing energy consumption. This work provides a reliable technical solution for safeguarding power grid operation and offers valuable references for tiny object detection in complex industrial scenarios.

Keywords: tiny bird detection; object location; bird repelling; heterogeneous binocular vision; transformer substation

1. Introduction

Substations are the core facilities that ensure the safe and stable operation of power grids. Most substations are located in suburban areas with well-preserved ecosystems, where wildlife activities are frequent. Among them, bird behaviors such as nesting, perching, and excretion pose severe threats to the normal operation of substation equipment, easily triggering faults such as short circuits and insulation failures, which directly undermine the reliability of the power grid supply [1]. Therefore, achieving efficient detection and repulsion of birds in substations is crucial to maintain the stable operation of substations. Traditional manual bird repulsion methods are not only high-cost and low-efficiency but also pose potential safety risks to personnel due to the high-voltage environment of substations, making them increasingly unable to meet practical demands. At present, substations mostly adopt automated bird repulsion devices such as acoustic [2], optical [3], and laser [4] devices. Although these devices have improved operational safety, they generally lack bird detection and positioning capabilities and need to operate continuously. This not only results in massive waste of energy and equipment resources but also introduces new safety hazards due to mechanical wear and energy radiation generated during long-term operation.

The rapid advancement of deep learning techniques has driven the widespread deployment of object detection technology. By leveraging object detection to acquire bird positions and trigger bird repulsion devices, proactive and precise bird repulsion can be achieved, which significantly improves the success rate of bird deterrence. Early object detection research primarily focused on large-scale targets. However, tiny targets pose extreme detection challenges due to their scarce feature information [5, 6], thus making tiny object detection a growing research hotspot in the field of object detection.

Nevertheless, existing achievements in tiny object detection are mostly confined to specific objects or general scenarios, exhibiting poor adaptability for bird detection in substation environments. The core bottleneck lies in the unique complexity of substation scenarios, and the associated challenges lack quantitative support and targeted research validation:

1) Background interference. Substations are characterized by a high proportion of interfering elements such as weeds, shadows, and equipment supports [7]. According to on-site statistics, within typical monitoring fields of view in substations, the pixel proportion of such interferences can reach 35%–45%. Moreover, their morphology and size are highly similar to those of tiny birds, which easily leads to false detections. Existing general tiny object detection methods suffer from high false detection rates under such complex background conditions, and no studies have yet optimized detection algorithms for background interference in substations.

2) Thermal radiation interference in infrared detection. Pseudo-target regions formed by thermal radiation from substation equipment [8] in infrared images increase the missed detection rate of traditional infrared tiny object detection algorithms. Most existing research on infrared object detection targets low-temperature backgrounds in natural scenarios, failing to account for thermal interference from high-temperature industrial equipment.

3) Lack of specialized datasets. The specificity of substation scenarios has resulted in the absence of public datasets for tiny bird targets in substations. Current studies mostly adopt general bird datasets for model training; however, the background distribution and target scales of general datasets differ significantly from those of substation scenarios. This leads to low detection accuracy of models in

actual substation environments, and no studies in existing literature have addressed dataset construction and adaptability research for bird detection in substations.

To address the aforementioned challenges, this paper develops a tiny bird detection and localization system integrated with a dual-modal detection algorithm. The framework is implemented through three core steps. First, a large number of bird scene images in substations are collected using infrared and visible-light cameras mounted on bird repulsion devices, and a self-constructed dataset is established. Second, a differentiated algorithm is adopted to achieve accurate bird detection under varying illumination and environmental conditions. At last, bird spatial localization is completed based on a heterogeneous binocular module, which drives the device to implement precise bird repulsion. The specific contributions of this paper are summarized as follows:

- A system integrating tiny bird detection and localization is constructed for substation scenarios, which synergizes dual-modal detection and binocular distance measurement. This system overcomes the limitations of single-modal detection and achieves targeted laser repulsion.
- For well-lit scenarios, a two-stage contextual information enhancement network is proposed for visible-light images. Drawing on the secondary feature extraction mechanism of DetectorRS [9], the network collaboratively exploits multiscale contextual information through convolution, dilated convolution, and linear layers. It combines channel attention to optimize localization accuracy and designs a dedicated training supervision strategy for tiny targets. This network effectively suppresses background interference and achieves an overall average precision (AP) of 59.8% on the self-constructed visible dataset (VSTBD), outperforming state-of-the-art methods by 2.5%.
- For low-light or occluded scenarios, a multiscale contextual feature enhancement network is proposed for infrared images. The network employs multibranch cross-layer feature fusion to retain tiny target contours, combines transformer and multisize convolution to model long-range feature dependencies, and implements multiscale prediction supervision. On the self-constructed infrared dataset (IRTBD), the network achieves a pixel accuracy ($PixAcc$) of 81.24% and an intersection over union (IoU) of 65.82%.
- A large-scale specialized dataset for substation tiny bird detection is constructed, including 20,000 visible images and 15,000 infrared images. The dataset covers diverse weather conditions, time periods, and substation environments, providing scenario-adapted training data for model generalization.

The subsequent structure of this paper is organized as follows: Section 2 reviews relevant works on tiny object detection in visible and infrared images, respectively; Section 3 introduces the overall system framework and elaborates on the proposed detection algorithms as well as the heterogeneous binocular detection and localization method; Section 4 presents the experimental results and conducts in depth analysis; Section 5 summarizes the research achievements and prospects future work.

2. Related work

2.1. Tiny object detection for visible images

Current research on tiny object detection mainly focuses on following technical directions to improve the network's ability to capture and represent tiny target features, thereby enhancing

detection performance: advanced convolution module integration, contextual information mining, multiscale feature fusion, and targeted training optimization.

Yue et al. [10] integrated depthwise separable convolution and channel shuffle modules to explore the deep-level features of the network, facilitate the fusion of local details and channel information, and enhance the network's capability of capturing tiny targets. Similarly, Mahaur et al. [11] improved YOLOv5 by replacing standard convolution with depthwise separable convolution, which achieved a balance between tiny object detection accuracy and overall efficiency. Contextual information mining is a mainstream strategy to make up for the lack of effective features of tiny targets by integrating contextual cues around the target. Zhao et al. [12] decoupled the contextual information of scenes by constructing a dedicated subnetwork for scene classification, which effectively integrated the scene context information around dense tiny targets and improved the detection performance for dense tiny targets. Xiao et al. [13] combined multiscale dilated convolution with channel-spatial conflict elimination to enhance contextual information, providing more abundant contextual support for tiny object detection. In addition, Huang et al. [14] constructed a feedback suppression attention module to suppress the background and large objects, which further strengthened the network's ability to focus on tiny targets by reducing the interference of irrelevant contextual information.

Multiscale feature fusion aims to solve the problem of insufficient feature expression caused by the large scale variation of tiny targets. By fusing features of different scales, the network can obtain more comprehensive and high-quality feature representations. Liu et al. [15] proposed a novel denoised feature pyramid; by introducing regularization operations, the pyramid eliminates noise from multiscale features and focuses on tiny targets via the self-attention mechanism. Leng et al. [16] proposed a bidirectional feature fusion network to transfer deep semantic features and shallow detailed features mutually, strengthening internal object features through contextual relationships. Ma et al. [17] proposed a multilevel weighted depth-aware network that fuses high-level and low-level feature maps and performs weighted merging of multiscale features to obtain high-resolution enhanced features.

Targeted training optimization focuses on improving the network's attention to tiny targets during the training process. Xu et al. [18] alleviated the mismatch problem by dynamically modeling priors, label assignment, and object representation. Liu et al. [19] introduced a supervision mechanism for tiny targets in the network training process; through the supervision of tiny target regions, the feature representation of these regions was enhanced, thereby improving the detection accuracy of tiny targets. Christof et al. [20] took a different approach by leveraging video motion information to enhance tiny object detection in surveillance scenarios, which expands the application scope of tiny object detection methods to video sequences.

Despite the significant progress made by the aforementioned methods in improving tiny object detection performance, they still have limitations in adapting to complex and specific application scenarios. Specifically, most advanced convolution module and contextual mining methods lack targeted interference suppression strategies for scenarios with dense background clutter, leading to difficulty in effectively distinguishing tiny targets from the background. Existing multiscale feature fusion methods often adopt fixed fusion modes, which cannot dynamically adjust to the scale variation of tiny targets in complex scenes. Targeted training optimization methods are mostly designed for general tiny targets and lack customization for the unique characteristics of specific target types. To address these gaps, we need a context enhancement mechanism tailored to substation

backgrounds that can highlight tiny bird regions while suppressing scene-specific interference. Thus, this paper proposes a two-stage contextual information enhancement network that integrates dilated convolution and channel attention to dynamically mine multiscale context and suppress interference, with a dedicated training supervision strategy for substation tiny birds.

2.2. Tiny object detection for infrared images

Infrared tiny object detection is typically transformed into a pixel-wise classification task due to the binary mask annotation format of most datasets. Current mainstream architectures can be categorized into three types based on feature enhancement mechanisms: dense nested feature interaction, multiscale feature coordination, and global-local feature comparison. Dense nested interaction enhances feature representation by cascading and fusing features of different layers.

Li et al. [21] introduced dense connections into the U-Net architecture and designed a dense nested attention network, which applies channel and pixel attention layer-by-layer to strengthen cascaded dense features. Wu et al. [22] proposed a nested U-Net structure: a small U-Net serves as the basic feature extraction module and is nested into a larger U-Net layer-by-layer, with an interactive-cross attention module to transfer information between low-level and high-level features during encoding and decoding. Multiscale feature coordination addresses the scale variation of tiny targets by aggregating multiscale information. Zhang et al. [23], inspired by spatial pyramid pooling, performed multiscale sampling on high-level features, calculated non-local information for each scale separately, and aggregated these features for decoding prediction. Huang et al. [24] proposed a local similarity pyramid module that enhances nonlocal information in aggregated multiscale features to guide the fusion of shallow and deep features. Global-local feature comparison mines tiny targets by modeling long-range dependencies.

These infrared detection methods face prominent bottlenecks in substation scenarios. The core issue is the severe thermal radiation interference from substation equipment, which forms pseudo-targets in infrared images. This is an issue that has not been considered in existing studies focusing on natural low-temperature backgrounds. Dense nested and multiscale fusion methods primarily enhance target features but lack mechanisms to distinguish between real tiny birds and thermal radiation pseudo-targets, resulting in a high rate of missed detection. The transformer-based method models global dependencies but lacks dedicated optimization for small-scale thermal pseudo-target suppression. To address these issues, we need a multiscale feature enhancement strategy that can retain tiny bird thermal contours while suppressing equipment thermal interference. Hence, this paper proposes a multiscale contextual feature enhancement network that combines multibranch cross-layer fusion with multiscale prediction supervision, effectively suppressing thermal radiation interference through targeted feature optimization.

3. Methodology

3.1. Heterogeneous binocular measurement system

3.1.1. Overview of the proposed system

To realize reliable detection and high-precision positioning of tiny bird targets in complex transformer substation scenarios, monocular depth estimation, visible binocular vision, and infrared

binocular vision were comprehensively evaluated, and the heterogeneous binocular vision scheme was finally adopted. The core focus of this work is bird object detection, and the designed device is equipped with both visible-light and infrared cameras to cover diverse detection scenarios. Meanwhile, these two cameras are naturally combined into a heterogeneous binocular module to achieve distance measurement via parallax calculation, thus achieving the integration of dual-modal detection and spatial positioning. Monocular depth estimation relies on semantic features and prior scene information to infer target depth, but it is not applicable to tiny bird targets in substation scenes.

Visible binocular systems can achieve high-precision positioning in well-lit environments, but they become ineffective in low-light, foggy, or nighttime scenarios due to texture detail loss, leading to a high detection failure rate. In contrast, infrared binocular systems can detect targets via thermal features in weak light, but they are severely interfered with by thermal radiation from substation equipment. Moreover, both homogeneous schemes only support single-modal detection, exhibiting limited adaptability to the complex and variable substation environment. In comparison, the heterogeneous binocular system in this work has two core advantages:

1) It integrates dual-modal detection, realizing bird detection via visible-light cameras under good illumination and via infrared cameras when visible-light detection fails, covering a wider range of scenarios.

2) It leverages the two cameras to form a binocular module for parallax-based distance measurement, realizing the organic integration of detection and positioning without additional hardware costs, which is more suitable for the engineering application requirements of substation bird repulsion.

Based on the above scheme comparison and demand analysis, a heterogeneous binocular measurement system is developed for bird detection and repulsion. As shown in Figure 1, the system consists of a heterogeneous binocular module, which is composed of a visible camera and a thermal infrared camera, an embedded industrial computer, and a laser transmitter. This paper uses the software development kit (SDK) provided by the camera manufacturer to perform secondary development of the heterogeneous binocular vision system. The embedded industrial control computer is used to deploy software algorithms to perform real-time processing of captured bird images. The laser transmitter is used to repel the birds by shining on them. To avoid inflicting harm on the birds, a 532 nm green laser with a maximum power of 3 W is adopted in this device, and only simulates a "waving stick" motion when scaring off birds. The laser is not directed at the birds for extended periods. Existing literature [4] has proven that it will not harm birds.

Compared with existing bird repellent systems, this system has three key improvements:

1) The heterogeneous binocular architecture overcomes the low-light detection limitation of visible-only algorithms by integrating infrared imaging.

2) Calibrated binocular cameras and integrated detection-positioning algorithms improve the laser hit rate and repulsion success rate by fusing bird detection results with spatial distance information.

3) The embedded industrial computer is innovatively integrated into the pan-tilt-zoom (PTZ) unit and powered directly by the PTZ unit, thus reducing the system complexity and the associated deployment costs.

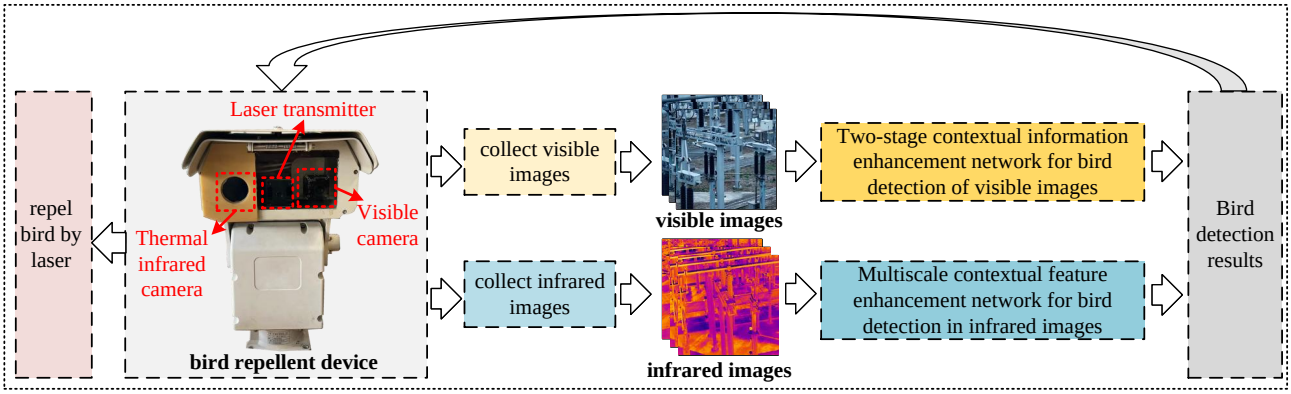
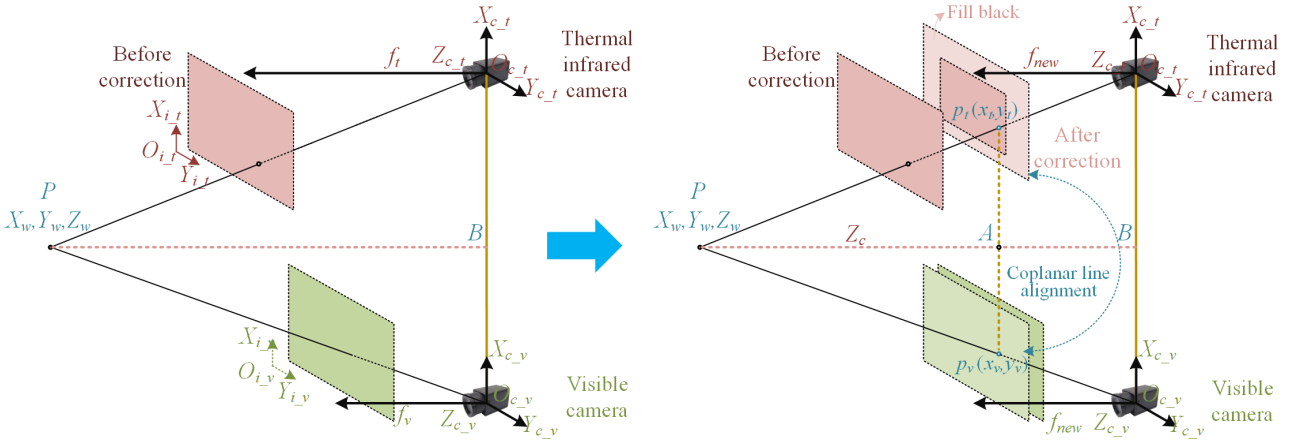


Figure 1. The bird repellent device proposed by the paper. Visible and infrared images are acquired by the visible light camera and thermal infrared camera in the designed device, respectively. The algorithms proposed in Sections 3.2 and 3.3 are then applied to detect and locate bird targets in the visible and infrared images. The obtained location information is transmitted to the embedded industrial computer to control the movement of the pan-tilt unit and the activation of the laser transmitter. The pan-tilt unit drives the laser to simulate a "waving stick" motion, thus scaring off the bird targets.

3.1.2. Calibration of heterogeneous binocular module and localization method

Different from the traditional binocular vision model, the heterogeneous binocular vision model requires the alignment of the imaging plane before stereo matching because the focal lengths of the two cameras differ. The simplified heterogeneous binocular stereo vision system is shown in Figure 2.



(a) Heterogeneous binocular vision model before correction

(b) Heterogeneous binocular vision model after correction

Figure 2. Simplified heterogeneous binocular stereo vision system.

The heterogeneous binocular module not only supports dual-modal bird detection but also realizes spatial positioning of bird targets through camera calibration and parallax calculation, providing precise

distance information for the laser repulsion. Assuming that the coordinate system of the visible camera is ($O_{c\_v} - X_{c\_v}\ Y_{c\_v}Z_{c\_v}$), the coordinate system of the thermal infrared camera is ($O_{c\_t} - X_{c\_t}\ Y_{c\_t}Z_{c\_t}$), and the world coordinate system is ($O_W - X_W Y_W Z_W$). For a point $P(X_W, Y_W, Z_W)$ in the world coordinate system, the corresponding point $P'(X_c, Y_c, Z_c)$ can be found in the camera coordinate system. The points of the two coordinate systems can be calculated by the rotation and translation rigid transformation matrix. The infrared camera coordinate system can be associated with the visible camera coordinate system with the help of the world coordinate system, so that the points in the two coordinate systems can be converted to each other:

$$\begin{bmatrix} X_{c\_v} \\ Y_{c\_v} \\ Z_{c\_v} \\ 1 \end{bmatrix} = \begin{bmatrix} R_v R_t^{-1} & T_v - R_v R_t^{-1} T_t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_{c\_t} \\ Y_{c\_t} \\ Z_{c\_t} \\ 1 \end{bmatrix} = \begin{bmatrix} R_{vt} & T_{vt} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_{c\_t} \\ Y_{c\_t} \\ Z_{c\_t} \\ 1 \end{bmatrix}, \tag{3.1}$$

where $R_{vt} = R_v R_t^{-1}, T_{vt} = T_v - R_v R_t^{-1} T_t$ are the rotation matrix and translation matrix that convert the infrared camera coordinate system to the visible camera coordinate system. Therefore, the two images can be converted to the same resolution. Then, the camera parameters are calibrated using Zhang's calibration method [25]. Combined with the camera parameters, the heterogeneous binocular images are stereo rectified. Finally, the depth value of point P is calculated based on the principle of similar triangles.

The calibrated intrinsic and extrinsic parameters of the heterogeneous binocular module lay the foundation for spatial localization. Subsequent Sections 3.2 and 3.3 will detail the dual-modal detection algorithms to obtain 2D bounding box coordinates of tiny bird targets. The spatial distance and position of targets will be calculated by combining these coordinates with the calibrated parameters and parallax principle.

For scenarios where the target is successfully detected by one modality but not the other, this study estimates the virtual coordinates of the target in the coordinate system of the non-detecting camera using the precalibrated relative pose parameters (rotation matrix R_{vt}, translation vector T_{vt}) of the heterogeneous binocular cameras, and completes spatial localization by combining these virtual coordinates with the real coordinates from the detecting camera. Meanwhile, false alarms in single-modal detection are suppressed through matching verification between target features and a feature library, ensuring the robustness of the system in complex scenarios.

3.2. Two-stage contextual information enhancement network for bird detection in visible images

The overall framework of the proposed two-stage contextual information enhancement network is described in this section. As shown in Figure 3, multiscale features are extracted twice via the pre-trained ResNet-50 to address the issue of tiny target omission in single-stage feature extraction. The first stage introduces the context information extraction and fusion (CIEF) module, which discriminates pixel-level differences between local neighborhoods to highlight salient tiny target regions, guiding the second stage to focus on potential bird targets. The high-level feature enhancement (HFE) module is then applied to strengthen localization cues of bird targets in high-level features, while the feature fusion pyramid network (FFPN) fuses cross-stage features and integrates spatial attention (SA) to suppress background noise and refine target positioning. Parallel

detection heads are employed to fully exploit the potential of multilevel features, and the normalized Wasserstein distance (NWD) [26] is embedded to ensure high detection precision.

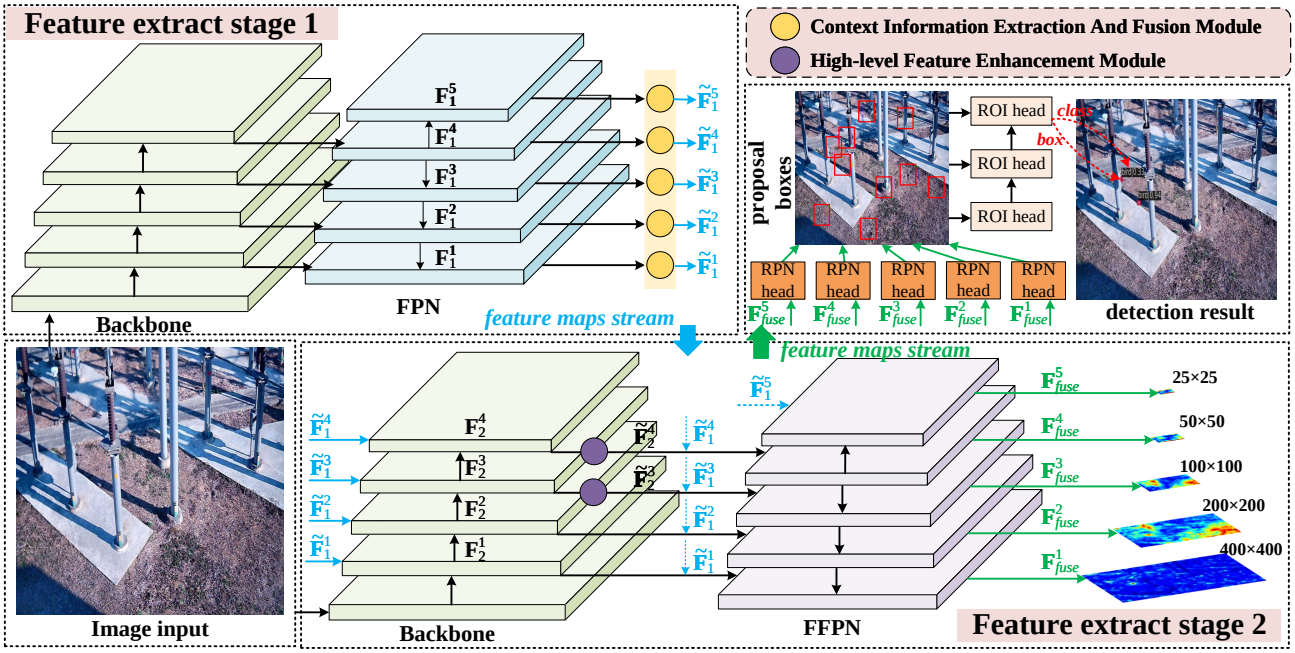


Figure 3. Architectural overview of the two-stage contextual information enhancement network for bird detection in visible images. In the first stage, all locally salient regions in the image are highlighted using the proposed context information extraction and fusion (CIEF) module (see Section 3.2.1). In the second stage, these salient regions are further enhanced, and the localization cues of tiny targets is strengthened by integrating the proposed high-level feature enhancement module (see Section 3.2.2). Subsequently, the features extracted from the two stages are fused in the feature fusion pyramid network (see Section 3.2.3), and the spatial localization of bird targets is further refined via spatial attention. Finally, the detection results of tiny birds are obtained via multiple detection heads.

3.2.1. Context information extraction and fusion

Most existing object detection methods use the information of the entire image to search objects pixel-by-pixel. This strategy performs well for large-scale object detection, yet incurs substantial computing resource wastage for tiny target scenarios. Birds usually only occupy a very small area of the entire high-resolution image. Unlike traditional context enhancement modules, which either mine global context in a single-stage manner or adopt fixed receptive fields for feature fusion and thus fail to suppress the background interference in substation scenes and easily confuse tiny birds with weeds or equipment supports, this paper constructs a context information extraction and fusion module (CIEF), as shown in Figure 4, which innovates in the following key aspects. First, it abandons the global image traversal strategy and focuses on local-neighbor feature interactions to avoid resource waste on irrelevant background regions. Second, it decouples local feature extraction and far-neighbor context modeling in parallel, rather than using a single receptive field to capture mixed information.

Finally, it integrates global context weighting via adaptive pooling and multilayer perceptron (MLP) to highlight true tiny bird regions while suppressing spurious salient regions induced by cluttered backgrounds in substations. This module uses local feature information to enhance the locally salient areas of tiny targets throughout the image, making it easier for the network to focus on tiny targets in images and improve the accuracy of bird detection.

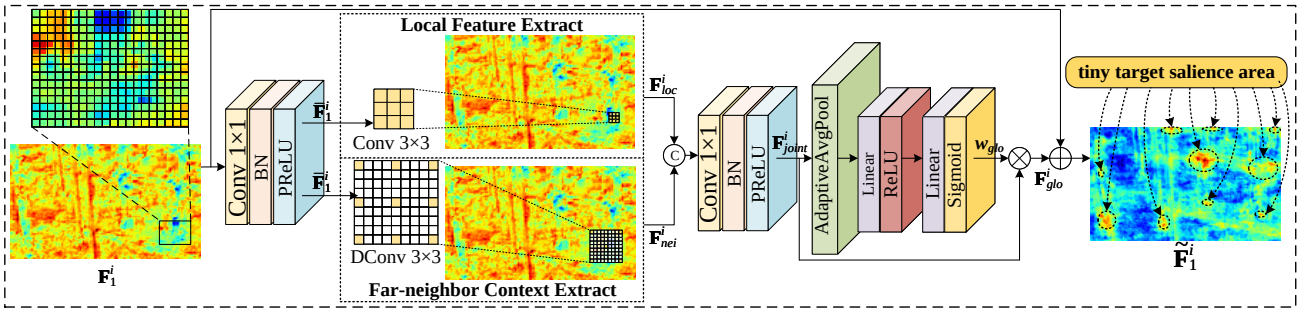


Figure 4. Context information extraction and fusion module.

At first, a convolutional operation is applied to the features extracted by ResNet-50 and the FPN. A local feature extractor and a long-neighbourhood context feature extractor are constructed for parallel operation:

$$\overline{\mathbf{F}}_1^i = f_{CBPR}^{1\times1}\left(\mathbf{F}_1^i\right),\ \mathbf{F}_{loc}^i = f_{loc}\left(\overline{\mathbf{F}}_1^i\right),\ \mathbf{F}_{nei}^i = f_{nei}\left(\overline{\mathbf{F}}_1^i\right), \tag{3.2}$$

where $f_{CBPR}^{1\times1}(\cdot)$ is implemented sequentially as a convolution operation with the size of 1×1, a batch normalization (BN) and a parametric rectified linear unit (PReLU). $f_{loc}(\cdot)$ denotes the local feature extractor, and $f_{nei}(\cdot)$ denotes the far-neighbor context extractor, which are instantiated as a 3×3 standard convolutional layer and a 3×3 atrous convolutional layer, respectively. $\mathbf{F}_1^i \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times C}$ denotes the ith layer input feature, $(i = 1, 2, 3, 4, 5)$, output from the corresponding layer of FPN. H, W, and C represent the height, width, and the number of image channels, respectively. $\overline{\mathbf{F}}_1^i \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times C}$ denotes the feature which is enhanced by $f_{CBPR}^{1\times1}(\cdot)$. $\mathbf{F}_{loc}^i \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times\frac{C}{2}}$ and $\mathbf{F}_{nei}^i \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times\frac{C}{2}}$ are the local feature information and the far-neighbor context information, respectively. Through the processing of two feature extractors, the network can highlight the features of tiny targets by only utilizing the feature information around the pixels. Then, $\mathbf{F}_{loc}^i$ and $\mathbf{F}_{nei}^i$ are sent to aggregate the features, and the aggregated feature is used to calculate the global context information:

$$\mathbf{F}_{joint}^i = f_{CBPR}^{1\times1}\left(\left[\mathbf{F}_{loc}^i, \mathbf{F}_{nei}^i\right]\right), w_{glo} = f_{MLP}\left(f_{Adap}\left(\mathbf{F}_{joint}^i\right)\right), \tag{3.3}$$

where $[\cdot,\cdot]$ denotes the concatenation operation. $\mathbf{F}_{joint}^i \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times C}$ denotes the aggregated feature. $f_{Adap}(\cdot)$ denotes the adaptive average pooling layer. $f_{MLP}(\cdot)$ denotes the multilayer perceptron, which consists of a linear layer followed by the rectified linear unit (ReLU) activation function and a linear layer followed by a sigmoid activation function. $w_{glo} \in \mathbb{R}^{1\times1\times C}$ denotes the weight vector of global

context information, which is calculated by $f_{MLP}(\cdot)$. The w_{glo} is used to weight $\mathbf{F}^i_{joint}$ in the way of skip connection, so as to emphasize useful features. Finally, residual learning is employed to help the network learn more complex features and improve the gradient back-propagation during training by connecting the input of the module with the output of the global operator, so as to obtain the feature $\tilde{\mathbf{F}}^i_1$, which contains the highlighted area of the tiny targets:

$$\mathbf{F}^i_{glo} = \mathbf{F}^i_{joint} \otimes w_{glo}, \tilde{\mathbf{F}}^i_1 = \mathbf{F}^i_1 \oplus \mathbf{F}^i_{glo}, \tag{3.4}$$

where $\mathbf{F}^i_{glo} \in \mathbb{R}^{\frac{H}{2^i} \times \frac{W}{2^i} \times C}$ denotes the weighted feature. $\otimes$ and $\oplus$ denote the multiplication in the channel dimension and addition by pixel, respectively. Then, $\tilde{\mathbf{F}}^i_1$ is added into the corresponding layers of the second backbone to obtain the second extracted features $\mathbf{F}^i_2$.

This module targets the pain point that tiny bird targets are easily obscured by background clutter in substation scenes. First, mark all suspected tiny target regions in the image by simultaneously extracting local features and far-neighbor context features, then assign higher weights to truly salient regions via global context modeling. This allows the network to focus on potential bird targets and avoid missing tiny targets due to background interference.

3.2.2. High-level feature enhancement

Since the features output by the CIEF only highlight the areas of tiny targets in the image, other areas of non-bird tiny targets in the image are also enhanced, which thus interferes with the subsequent detection of bird targets. The channel attention mechanism allows the network to adaptively assign weights to different channels according to their importance, thus enhancing useful information and suppressing useless information. Compared with low-level features, high-level features with smaller spatial dimensions contain richer discriminative information for bird targets. To enhance the discriminative capability of bird-related feature information, this paper constructs a high-level feature enhancement module (HFE) by combining the channel attention mechanism to enhance the feature information of birds in the high-level features. The details of the module are shown in Figure 5.

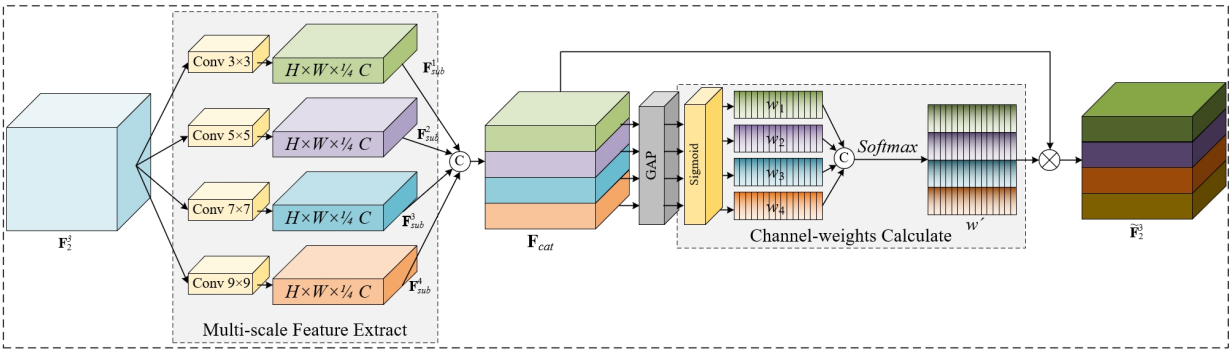


Figure 5. High-level feature enhancement module.

Take feature $\mathbf{F}^3_2 \in \mathbb{R}^{\frac{H}{2^3} \times \frac{W}{2^3} \times C}$ as an example. First, the feature map is divided into four branches that are convolved respectively. Then, the feature maps are compressed in the channel dimension so that

the channel number of each output feature map is $\frac{C}{4}$. The spatial information with different scales on each channel-wise feature map can be effectively extracted by squeezing the channel dimension of the input. The four subchannel feature maps are concatenated after squeezing:

$$\mathbf{F}_{sub}^{j} = f_{conv}^{j}\left(\mathbf{F}_{2}^{3}\right), \mathbf{F}_{cat} = \left[\mathbf{F}_{sub}^{1}, \mathbf{F}_{sub}^{2}, \mathbf{F}_{sub}^{3}, \mathbf{F}_{sub}^{4}\right], \tag{3.5}$$

where $f_{conv}^{j}(\cdot)$ denotes the convolution operator of which kernel size is $(2j+1) \times (2j+1)$, $j = \{1, 2, 3, 4\}$, so as to ensure that the size of the output feature map is consistent with the input feature map in width and height. $\mathbf{F}_{sub}^{j} \in \mathbb{R}^{\frac{H}{2^3} \times \frac{W}{2^3} \times \frac{C}{4}}$ denotes the jth subchannel feature map. $\mathbf{F}_{cat} \in \mathbb{R}^{\frac{H}{2^3} \times \frac{W}{2^3} \times C}$ denotes the concatenated feature. In order to realize the interaction between local and global channel attention, the four sub-channel features are used to calculate the channel attention weight vectors and normalized by Softmax function after concatenating:

$$w_{j}^{'} = \sigma\left(GAP\left(\mathbf{F}_{sub}^{j}\right)\right), w^{'} = f_{softmax}\left(\left[w_{1}^{'}, w_{2}^{'}, w_{3}^{'}, w_{4}^{'}\right]\right), \tag{3.6}$$

where $w_{j}^{'} \in \mathbb{R}^{1 \times 1 \times \frac{C}{4}}$ denotes the jth channel attention weight vector. σ denotes the Sigmoid function. $w^{'}$ denotes the normalized weight vector. Then $\mathbf{F}_{cat}$ is multiplied with $w^{'}$ to obtain the enhanced feature $\tilde{\mathbf{F}}_{2}^{3} = \mathbf{F}_{cat} \otimes w^{'}$.

The core goal of HFE module is to filter out non-bird interference and amplify bird-specific features following the marking of suspected regions by the CIEF module. It can be understood as splitting high-level features into multiscale convolution branches to adapt to birds of varying postures, then using channel attention to assign higher weights to channels containing bird features and suppress channels dominated by background clutter.

3.2.3. Feature fusion pyramid network

The areas suspected of being a bird in the feature map are removed by the HFE. However, there is still useless background information in the feature map. The SA mechanism [27] can locate the key information in the feature map and enhance it with the spatial weights. Therefore, this paper designs a feature fusion pyramid network module (FFPN), which is combined with the spatial attention mechanism to further locate the feature information of the bird in the feature map and suppress useless background information. The details of the module are shown in Figure 6. By alternately using the key information of the feature maps from different stages, the features extracted in the first and second stages are fused. Meanwhile, the SA weights are calculated by the SA module. The features of each stage are enhanced based on the SA weights of each feature.

First, the feature from the second stage, at a higher level, is upsampled to the same size as the feature from the lower level. Then, the two features are added pixel by pixel. The added result and the same-scale feature from the first stage are used to calculate the feature $\mathbf{F}_{m}^{1}$, which contains feature information from two stages:

$$\hat{\mathbf{F}}_{2}^{1} = \mathbf{F}_{2}^{1} \oplus f_{up}\left(\mathbf{F}_{2}^{2}\right), \mathbf{F}_{m}^{1} = f_{CBPR}^{1\times1}\left(\tilde{\mathbf{F}}_{1}^{1}\right) \otimes f_{CBPR}^{1\times1}\left(\hat{\mathbf{F}}_{2}^{1}\right), \tag{3.7}$$

where $\mathbf{F}_2^1 \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times C}$ and $\mathbf{F}_2^2 \in \mathbb{R}^{\frac{H}{2^2}\times\frac{W}{2^2}\times C}$ denote the adjacent low-level feature and high-level feature in the second stage, respectively. $\tilde{\mathbf{F}}_1^1 \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times C}$ denotes the feature of the first layer enhanced by the CIEF module. $\mathbf{F}_m^1 \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times C}$ denotes the feature map, which contains abundant local information from the feature maps in two steps. Then, $\mathbf{F}_m^1$ is added to $\tilde{\mathbf{F}}_1^1$ and $\hat{\mathbf{F}}_2^1$ respectively, in order to enhance the local features in the features of each stage:

$$\mathbf{F}_{1\_m}^1 = \tilde{\mathbf{F}}_1^1 \oplus \mathbf{F}_m^1, \mathbf{F}_{2\_m}^1 = \hat{\mathbf{F}}_2^1 \oplus \mathbf{F}_m^1. \tag{3.8}$$

Then, a dilated convolution layer is used after concatenating $\mathbf{F}_{1\_m}^1$ and $\mathbf{F}_{2\_m}^1$, which extracts the global features of the feature maps. The dilated convolution layer is followed by $f_{CBPR}^{1\times1}(\cdot)$. Additionally, the SA mechanism is used to calculate the SA weight of bird features. Combined with the SA weight, the network can further fuse the local and global information of the features:

$$\mathbf{F}_{fuse}^1 = \left[\tilde{\mathbf{F}}_1^1 \otimes \mathcal{S}\left(\hat{\mathbf{F}}_2^1\right), \hat{\mathbf{F}}_2^1 \otimes \mathcal{S}\left(\tilde{\mathbf{F}}_1^1\right), f_{CBPR}^{1\times1}\left(\mathcal{D}\left(\left[\mathbf{F}_{1\_m}^1, \mathbf{F}_{2\_m}^1\right]\right)\right)\right], \tag{3.9}$$

where $\mathcal{S}(\cdot)$ denotes the SA mechanism, $\mathcal{D}(\cdot)$ denotes the dilated convolution, and $\mathbf{F}_{fuse}^1 \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times C}$ denotes the fused feature.

This module resolves the contradiction between shallow features that are rich in detail but weak in semantics and deep features that are strong in semantics but coarse in localization in two-stage extraction. Its core logic is to align and fuse features from the two stages to integrate fine-grained details and high-level semantics, then to use SA to focus on bird pixel positions while weakening background responses.

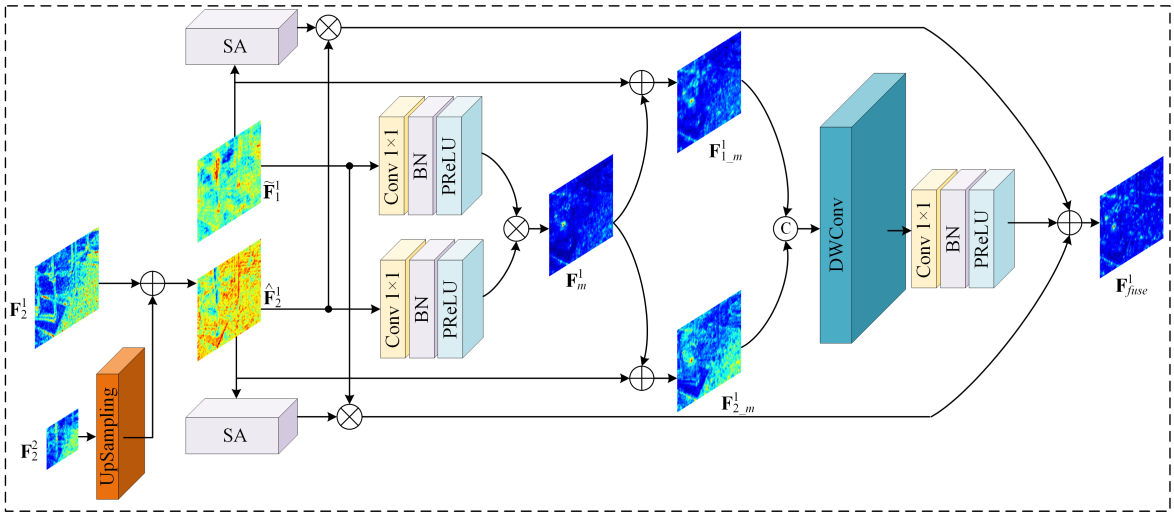


Figure 6. Feature fusion pyramid network.

3.2.4. Loss function

After extracting the features of the images, the features $\mathbf{F}_{fuse}^1 \sim \mathbf{F}_{fuse}^5$ are sent to the region proposal network (RPN) head and region of interest (RoI) head for classification prediction and bounding box regression. The cross-entropy loss function and smooth *L1* loss function are used to calculate the

classification loss and regression loss of the bounding box, respectively. Additionally, the strategy of assigning positive and negative samples based on IoU, which is adopted after the RPN head generates candidate boxes, suffers from inherent limitations due to the small size of bird targets. Therefore, this paper adopts the normalized Wasserstein distance (NWD) and ranking-based assigning (RKA) optimization strategies [26] as the discrimination strategies for positive and negative samples.

3.3. Multiscale contextual feature enhancement network for bird detection in infrared images

Due to the limitations of bird detection algorithms in visible images, such as lighting and weather conditions, the infrared imaging is introduced to address it. Infrared sensors can detect the thermal radiation emitted by birds, aiding in the detection of bird targets in complex environments. Therefore, the paper proposes a multiscale contextual feature enhancement network for bird detection in infrared images. Different from existing multibranch transformer networks that focus on global dependency modeling but ignore scene-specific interference suppression, our network innovates in branch design and feature interaction mechanisms to tackle the core pain point of thermal radiation pseudo-target interference in substation scenarios. As shown in Figure 7, the proposed network consists of three modules. First, the multiscale feature extraction module is designed to extract multiscale features based on the feature interactions between the backbone and branches at different scales. Unlike the branch design in conventional multibranch transformer networks, we configure three differentiated branches to balance deep semantic feature mining and shallow target contour retention, which avoids the loss of tiny bird thermal features and the confusion with substation equipment thermal artifacts. After that, the extracted features are transmitted into contextual information extraction module designed to suppress background interference and thereby enhance the salience of bird targets. Finally, multiscale prediction supervision facilitates the parameter learning during network training and generates prediction results during the network inference, which further compensates for the lack of multilevel feature supervision in existing transformer-based methods.

3.3.1. Multiscale feature extraction

The features are easily overwhelmed or disrupted by background interference due to the diminutive size of tiny bird targets. To address this issue, a multiscale feature extraction module based on the critical frameworks of the Xception network [28] and dual path network [29] is proposed. Its structure is illustrated in Figure 7. Specifically, multiple layers of convolution, followed by max pooling, are employed to progressively reduce feature scales and extract multiscale features $\mathbf{F}_i \in \mathbb{R}^{C_i \times H_i \times W_i}$, where $i = 0, 1, 2, 3, 4, 5$. The input size is denoted as $H \times W$, so H_i is $\frac{H}{2^i}$, W_i is $\frac{W}{2^i}$, and $C_i \in \{64, 128, 256, 512, 512, 512\}$. In addition to the feature extraction trunk, a detail-preservation branch and a multiscale feature fusion branch are designed to supplement the trunk features so as to obtain multiscale features.

The detail-preservation branch. The selection of the detail-preserving branch uses shallow features $\mathbf{P}_1$ as input, which contains rich detail information to supplement deep features. The saliency of tiny target features can be further stimulated and adapted to the size of main features by recoding the features $\mathbf{P}'_1 = f^{1\times1}_{CBR}(\mathbf{P}_1)$, where $\mathbf{P}'_1$ is the recoding feature. $f^{1\times1}_{CBR}(\cdot)$ is implemented sequentially as a convolution operation with a size of 1×1, a BN, and a ReLU. To efficiently and simply supplement the feature information, $\mathbf{P}'_1$ is directly down-sampled and summed as the input feature for the next stage

of coding. Furthermore, $\mathbf{P}_1'$ is recoded twice and merged with the features from the following two stages respectively.

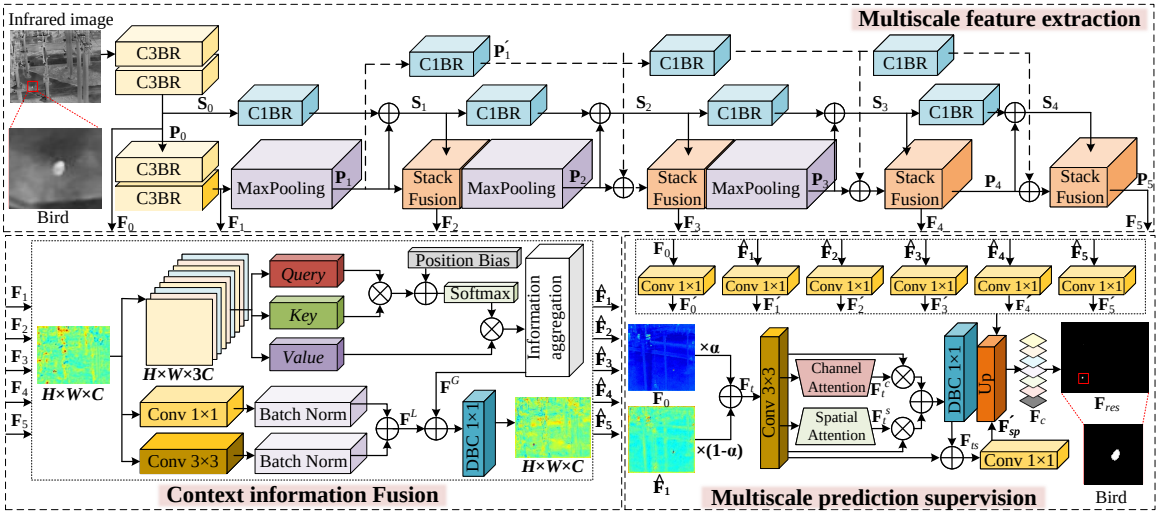


Figure 7. Architectural overview of the multiscale contextual feature enhancement network for bird detection in infrared images. First, multiscale features of different branches are extracted via the constructed multiscale feature extraction module (see Section 3.3.1). In the contextual information extraction module (see Section 3.3.2), background suppression is performed on the multiscale features to highlight bird saliency. Finally, the detection results are further optimized by integrating multiscale prediction supervision (see Section 3.3.3).

The multiscale feature fusion branch. A complex background, abnormal weather, and other conditions can lead to a reduction in the contrast between tiny bird targets and the background, which can cause the loss of bird targets information in deep features. To maintain the salience and balance the spatial positioning and semantic information of bird targets during the process of deep feature coding, multilevel features are gradually merged in the multiscale feature fusion branch. First, shallow feature $\mathbf{P}_0$ is taken as the input feature. The features of each scale need to be processed to optimize the representation of the feature map. The extracted features from the backbone are combined with the corresponding features from the multiscale feature fusion branch to obtain fused features, $\mathbf{S}_j = \mathbf{P}_j + f_{CBR}^{1\times1}\left(\mathbf{S}_{j-1}\right)$, where $j = 1, 2, 3, 4$ represents the stages of feature fusion. Correspondingly, the branch feeds the fused features back into the backbone branch. To mitigate the inherent information conflict between features of different scales, stack fusion is applied for integrating features. The stack fusion is composed of k fusion blocks, where k represents the number of fusion blocks. In this paper, k is set to 3. In each fusion block, the branch feature $\mathbf{S}_j$ is fused with the backbone feature after $f_{CBR}^{3\times3}(\cdot)$:

$$\mathbf{P}_j^1 = \left[f_{CBR}^{3\times3}\left(\mathbf{S}_j\right) + f_{CBR}^{3\times3}\left(f_{CBR}^{3\times3}\left(\mathbf{P}_j\right)\right)\right] \times \lambda, \tag{3.10}$$

where $f_{CBR}^{3\times3}(\cdot)$ is implemented sequentially as a convolution operation with a size of 3×3, followed by a BN and a ReLU, λ serves as a constant variable of 0.5 to maintain the balance of eigenvalues

and prevent fusion ratio imbalance due to accumulation. $\mathbf{P}_j^1$ represents the feature after the first fusion block within the stack fusion process. After three feature fusion blocks, the output $\mathbf{F}_{j+1}$ is generated. Additionally, $\mathbf{F}_0 = \mathbf{P}_0 = \mathbf{S}_0$ at the beginning.

This design addresses the pain point that bird thermal features in infrared images are easily masked by thermal radiation from high-temperature substation equipment. By using the backbone to extract deep semantic features, the detail-preservation branch to retain shallow target contours, and the multiscale fusion branch to integrate cross-level features and balance spatial localization and semantic representation. This design effectively distinguishes real bird thermal features from equipment thermal artifacts.

3.3.2. Contextual information extraction

To enhance the saliency of bird target features, a contextual information extraction module is designed to capture contextual information across multiscale features. The structure of this module is shown in Figure 7. This module combines convolution operations of different sizes with transformer operations to analyze relationships across various regions. For images with complex backgrounds, it helps suppress background noise, thereby highlighting the features of bird targets. The module is primarily divided into two parts for parallel processing. Taking the input features $\mathbf{F}_1 \in \mathbb{R}^{B\times C_1\times H_1\times W_1}$ with a batch size of B as an example, local contextual information is extracted in the local contextual information extraction branch:

$$\mathbf{F}_1^L = f_{BN}\left(f_{conv}^{1\times1}(\mathbf{F}_1)\right) + f_{BN}\left(f_{conv}^{3\times3}(\mathbf{F}_1)\right), \tag{3.11}$$

where $f_{BN}(\cdot)$ denotes batch normalization. $f_{conv}^{1\times1}(\cdot)$ and $f_{conv}^{3\times3}(\cdot)$ represent convolution operations with sizes of 1×1 and 3×3, respectively. During the process of global contextual information extraction, the dimensions of the input features are expanded to $\mathbb{R}^{B\times3C_1\times H_1\times W_1}$ through convolution operations. To improve the computational speed of self-attention, the feature map is divided into $\frac{H_1}{w}\times\frac{W_1}{w}$ windows, with each window having a size of w. Then, the feature is split into h heads in the channel dimension and further divided into three vectors $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V} \in \mathbb{R}^{\left(B\times\frac{H_1}{w}\times\frac{W_1}{w}\times h\right)\times(w\times w)\times\left(\frac{C_1}{h}\right)}$. The window-based multi-head self-attention (W-MSA) mechanism of the swin transformer [30] is utilized to integrate contextual information. An asymmetric pooling method is adopted to aggregate long-range information from adjacent windows. As shown in Figure 8, for a pixel $p_0^{(m,n)}$ located in the top-left window of the feature map, all feature values from $p_0^{(m,n)}$ to $p_1^{(m+w,n)}$ are aggregated through horizontal average pooling. This aims to explore the dependency between two horizontally adjacent windows:

$$p_0^{(m,n)} = \frac{\sum_{i=0}^{w-m-1} p_0^{(m+i,n)} + \sum_{j=0}^{m} p_1^{(m+w-j,n)}}{w}, \tag{3.12}$$

where $m \in 0, 1, ..., \frac{H_1}{w}$ and $n \in 0, 1, ..., \frac{W_1}{w}$. The pixel $p_0^{(m+i,n)}$ on the blue path has already computed self-attention with $p_0^{(m,n)}$, while the pixel $p_1^{(m+w-j,n)}$ on the purple path has computed self-attention with $p_1^{(m+w,n)}$. Thus, the average pooling described above enables the aggregation of contextual information horizontally. The dependencies between vertically adjacent windows are explored in the same manner:

$$\tilde{p}_0^{(m,n)} = \frac{\sum_{i=0}^{w-m-1} p_0^{(m,n+i)} + \sum_{j=0}^{m} p_2^{(m,n+w-j)}}{w}. \tag{3.13}$$

Finally, the results of horizontal and vertical pooling are summed to obtain the global contextual pooling information of $\mathbf{F}_1$, denoted as $\mathbf{F}_1^G$. After that, the local contextual information $\mathbf{F}_1^L$ and the global contextual information $\mathbf{F}_1^G$ are further fused to explore the salient features of the bird targets:

$$\mathbf{F}_1^G = p_0^{(m,n)} + \tilde{p}_0^{(m,n)}, \hat{\mathbf{F}}_1 = f_{DBC}^{1\times1}\left(\mathbf{F}_1^L \oplus \mathbf{F}_1^G\right), \tag{3.14}$$

where $f_{DBC}^{1\times1}(\cdot)$ contains a 1×1 deep convolution operation, a BN, and a convolution operation with a size of 1×1.

This module is to integrate global information to distinguish bird targets from background thermal noise in infrared images using convolutions to capture local thermal features of birds and transformer-based cross-window attention to integrate contextual information to filter out isolated thermal artifacts from equipment.

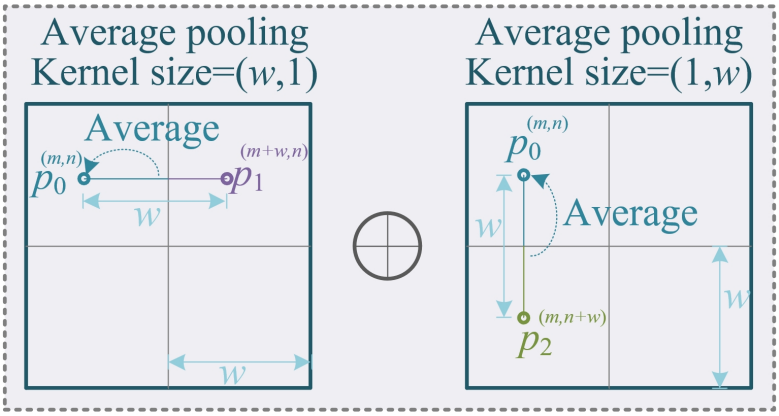


Figure 8. Window mapping and cross-window information aggregation.

3.3.3. Multiscale prediction supervision

Although shallow features have high resolution and fine local details, they are easily disturbed by background and the lack of spatial context guidance, which results in a large number of false positives in the classification results. Additionally, due to the small size of bird targets, it is difficult to extract sufficient semantic information from deep features, making the bird features prone to loss. As illustrated in Figure 7, we directly use the fused feature of shallow features, $\mathbf{F}_0$ and $\hat{\mathbf{F}}_1$, to compute the saliency map, thereby constraining the extraction of shallow features. To maintain lightweight computation in the network, the two features are simply fused based on learnable parameters α:

$$\mathbf{F}_t = \alpha \cdot \mathbf{F}_0 \oplus (1-\alpha) \cdot \hat{\mathbf{F}}_1. \tag{3.15}$$

The fused feature $\mathbf{F}_t$ is processed by channel attention and SA in parallel. The features in the pixel dimension are aggregated into a single-channel SA map $\mathbf{F}_t^s \in \mathbb{R}^{H\times W\times 1}$ by using a depth-wise separable convolution, in which H and W denote the spatial resolution of the feature map. Simultaneously, the spatial resolution of $\mathbf{F}_t$ is pooled to 1×1 and processed by two one-dimensional convolutions to aggregate channel information to obtain the channel attention map $\mathbf{F}_t^c \in \mathbb{R}^{1\times 1\times C}$:

$$\mathbf{F}_t^s = \delta\left(f_{DWC}^{3\times 3}(\mathbf{F}_t)\right), \mathbf{F}_t^c = \delta\left(f_{conv}^{1-D}\left(f_{conv}^{1-D}\left(f_{pool}^{avg}(\mathbf{F}_t)\right)\right)\right), \tag{3.16}$$

where $f_{DWC}^{3\times 3}(\cdot)$ denotes the depth-wise separable convolution with a kernel size of 3×3. $f_{conv}^{1-D}(\cdot)$ denotes the one-dimensional convolution. $f_{pool}^{avg}(\cdot)$ denotes average pooling. $\delta(\cdot)$ denotes sigmoid activation. $\mathbf{F}_t$ is multiplied by $\mathbf{F}_t^s$ and $\mathbf{F}_t^c$, respectively. The results are then summed with $\mathbf{F}_t$ in a residual manner to calculate the enhanced feature $\mathbf{F}_s$:

$$\mathbf{F}_s = f_{DBC}^{1\times 1}\left(\left(\mathbf{F}_t \otimes \mathbf{F}_t^s\right) \oplus \left(\mathbf{F}_t \otimes \mathbf{F}_t^c\right)\right) \oplus \mathbf{F}_t. \tag{3.17}$$

Furthermore, the low-level auxiliary prediction feature $\mathbf{F}_{sp}^{'}$ is obtained from the feature $\mathbf{F}_s$, and multiscale supervise features $\mathbf{F}_i^{'}$ are calculated:

$$\mathbf{F}_{sp}^{'} = f_{conv}^{1\times 1}\left(\mathbf{F}_s\right), \mathbf{F}_0^{'} = f_{conv}^{1\times 1}\left(\mathbf{F}_0\right), \mathbf{F}_i^{'} = f_{conv}^{1\times 1}\left(\hat{\mathbf{F}}_i\right), \tag{3.18}$$

where $i = 1, 2, 3, 4, 5$. The $\mathbf{F}_i^{'}$ and $\mathbf{F}_{sp}^{'}$ are upsampled to the size of the input image and concatenated along the channel dimension, denoted as $\mathbf{F}_c \in \mathbb{R}^{H\times W\times 7}$. The final prediction result $\mathbf{F}_{res}$ is then obtained through the convolution operation with a sigmoid activation function:

$$\mathbf{F}_c = \left[f_{up}\left(\mathbf{F}_0^{'}\right), ..., f_{up}\left(\mathbf{F}_5^{'}\right), f_{up}\left(\mathbf{F}_{sp}^{'}\right)\right], \mathbf{F}_{res} = \delta\left(f_{conv}^{1\times 1}\left(\mathbf{F}_c\right)\right), \tag{3.19}$$

where $f_{up}(\cdot)$ denotes the upsample function. This module targets the trade-off between shallow feature detail and deep feature semantics in infrared tiny object detection. Fusing shallow and deep features to retain both target contours and semantic information. Then, we use dual attention mechanisms to enhance target saliency and apply multiscale supervision to optimize feature learning at all levels. This ensures that no bird targets are missed due to feature loss in deep layers or background interference in shallow layers.

3.3.4. Loss function

In order to address the issue of class imbalance in binary classification tasks, the soft intersection over union ($SoftIoU$) loss function is employed:

$$\mathcal{L}_{SoftIoU}(P, M) = 1 - \frac{(P \bigcap M + smooth)}{(P \bigcup M + smooth)}, \tag{3.20}$$

where P and M denote the predicted result and the ground truth mask, respectively. *smooth* denotes the smooth factor, usually set to 1×10^{-7} to prevent zero in the union calculation. The overall loss function is calculated as follows:

$$\mathcal{L}_{SoftIoU} = \beta_1 \cdot \mathcal{L}_{SoftIoU}\left(\mathbf{F}_{res}, M\right) + \beta_2 \cdot \mathcal{L}_{SoftIoU}\left(\mathbf{F}'_{sp}, M\right) + \sum_{i=0}^{5} \gamma_i \cdot \mathcal{L}_{SoftIoU}\left(\mathbf{F}'_i, M\right), \tag{3.21}$$

where β_1, β_2, and γ_i are the weights corresponding to the different losses. Both of them are set to 1 to balance the effect of supervision in each scale.

4. Experiments

4.1. Dataset and implementation details

4.1.1. Datasets construction

To enhance the persuasiveness of the proposed algorithm, 20,000 visible images and 15,000 infrared images were collected under identical conditions, which were used to construct the visual tiny bird detection dataset (VSTBD) and the infrared tiny bird detection dataset (IRTBD), respectively, for experimental validation. Figure 9 presents example views of partial sample images from the datasets. These images were acquired from 8 independent devices, each equipped with both infrared and visible cameras and mounted on the rooftops of monitoring rooms across 4 distinct substation stations. Specifically, each substation contributed 25% of the visible images and 25% of the infrared images, thereby ensuring the spatial diversity of the datasets. Considering the impact of weather conditions, the datasets incorporate images captured under various meteorological scenarios, including rainy, foggy, and sunny days, with respective proportions of 15%, 10%, and 75%. Since images captured at night are not applicable for visible object detection, all images were collected during the time window from early morning to late evening. Additionally, data acquisition spanned multiple seasons to guarantee the temporal continuity of the dataset. To ensure the validity of annotations, a motion object detection algorithm was employed to analyze consecutive frame images and extract motion-related information within the frames. Subsequently, moving bird targets were screened via manual review and annotated using the Labelme tool. Upon completion of annotation, the labeling results were inspected to ensure annotation consistency. In terms of the number of bird instances, VSTBD contains 32,687 annotated bird targets, while IRTBD includes 21,542 instances. Given that this study focuses on bird detection rather than species-specific classification, no detailed categorization of bird species was performed. The target distribution and size distribution within VSTBD are illustrated in Figure 10. The aspect ratio of most targets in the images is less than 0.01. For compatibility with network training, all visible-light images were uniformly resized to 800×800 pixels, and infrared images were resized to 512×512 pixels. For the division of training and test sets, a stratified sampling strategy was adopted to preserve the distribution of weather conditions, time periods, and target densities between the two subsets, with 80% of the images allocated to the training set and the remaining 20% to the test set. To facilitate future research on bird object detection, the datasets will be made publicly available in the future.

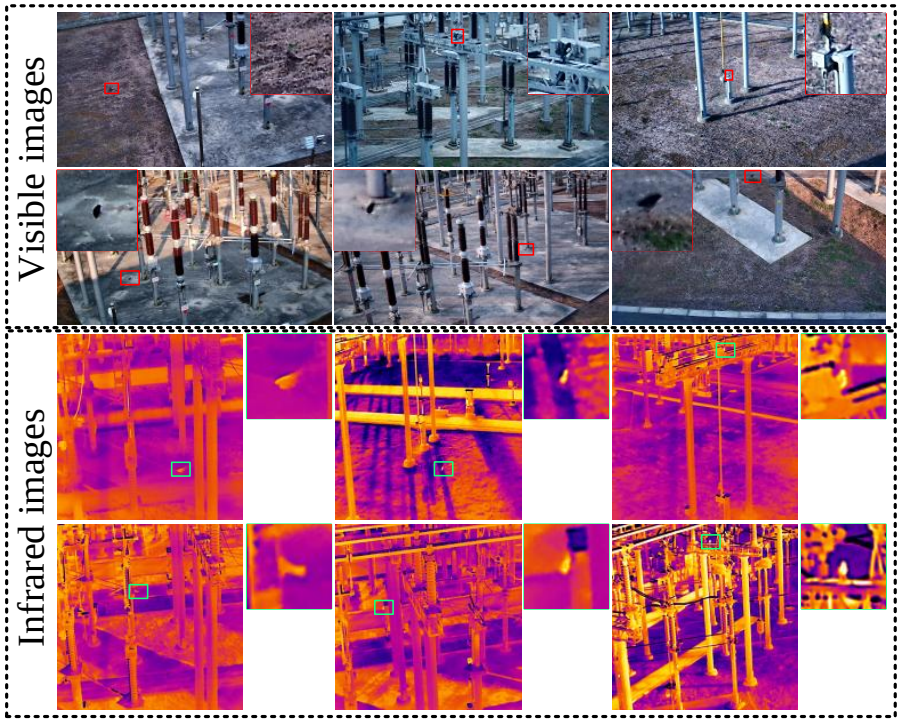


Figure 9. Dataset partial sample visualization.

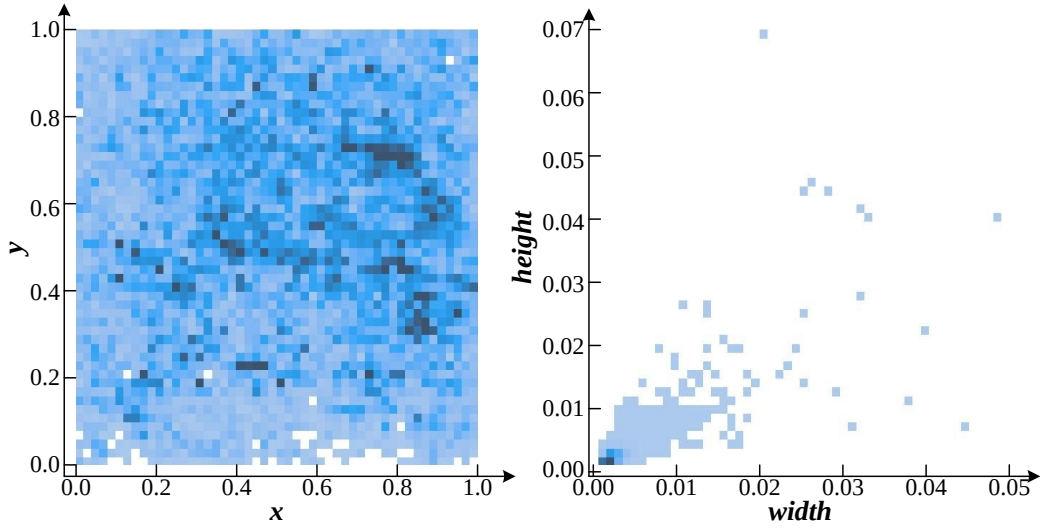


Figure 10. Analysis of distributions and sizes of boxes in visual tiny bird detection dataset.

4.1.2. Implementation details

All the experiments are conducted on a server equipped with an Intel Xeon 6226R 2.9GHz CPU, 18GB RAM, and two RTX 3090 24GB GPUs. The proposed detection algorithms are primarily implemented using Python and the deep learning library of Pytorch. In the train of the detection algorithm in visible images. The batchsize and the total number of training epochs are set to 4 and 12,

respectively. During the training process, the network parameters are optimized using the stochastic gradient descent algorithm with a momentum of 0.9 and a weight decay of 0.0001. The initial learning rate is set to 0.01 initially and is adjusted to 1% of its value at the 9th and 11th epoch. In the process of training and testing, the visible light image is interpolated to the resolution of 800×800. In the train of the detection algorithm in infrared images, the batchsize and the total number of training epochs are set to 32 and 400, respectively. The Adam optimizer is selected to optimize network parameters during training. The initial learning rate is set to 5×10^{-4} and reduced to 10% of its value at the epochs of 200 and 300. During the process of training, the input image will be randomly cut into blocks with the size of 256×256 to further increase the data richness, and the image size during the test will be uniformly scaled to 512×512.

Due to challenges such as small size, indistinct colors, and morphological variations, classical object detection algorithms cannot be directly applied to the tiny bird detection task in visible images. Therefore, this paper selects the faster region-based convolutional neural network method (Faster R-CNN) [31], the fully convolutional one-stage object detector (FCOS) [32], Cascade R-CNN [33], DetectorRS [9], the faster high resolution network (FasterHRNet) [34], RetinaNet [35], YOLOv8 [36], and YOLOX [37] object detection methods from the MMDetection framework as the comparative methods. Additionally, the receptive field based label assignment method (RLFA) [38], the dot distance method (DotD) [39], coarse-to-fine pipeline and feature imitation learning method (CFINet) [40], and NwdRka [26] are incorporated into the network training process. For the comparative experiment of tiny bird detection in infrared images, namely U-Net, asymmetric contextual modulation (ACM) [41], attentional local contrast network (ALCNet) [42], ISTDUNet [43], interior attention-aware network (IAANet) [44], receptive-field and direction-induced attention network (RDIAN) [45], dense nested attention network (DNANet) [21], and U-Net in U-Net (UIUNet) [22] are selected as comparative methods.

To eliminate the impact of experimental configuration discrepancies on model performance comparison, all baseline detectors and the proposed method were implemented based on the MMDetection framework. They were trained and tested on the VSTBD and IRTBD datasets, respectively, according to different task requirements. During the training phase, all models adopted consistent preprocessing pipelines, a unified input resolution, and identical hyperparameter settings. The network architecture of the proposed visible light image object detection and infrared image object detection algorithms is summarized in Tables 1 and 2. The specific training parameters for the visible and infrared image detection networks are summarized in Table 3.

Table 1. Network architecture overview of tiny bird object detection algorithm for visible images.

Network stage	Module	Feature Resolution	Channel Number
First-stage Feature Extraction	ResNet-50	800×800, 400×400, 200×200, 100×100, 50×50	3, 64, 256, 512, 1024
	FPN	400×400, 200×200, 100×100, 50×50, 25×25	64, 256, 512, 1024, 2048
	CIEF Module	400×400, 200×200, 100×100, 50×50, 25×25	64, 256, 512, 1024, 2048
Second-stage Feature Extraction	ResNet-50	800×800, 400×400, 200×200, 100×100, 50×50	3, 64, 256, 512, 1024
	HFE Module	100×100, 50×50	512, 1024
	FFPN Module	400×400, 200×200, 100×100, 50×50, 25×25	64, 256, 512, 1024, 2048
Detection Head	RPN+RoI Head	400×400, 200×200, 100×100, 50×50, 25×25	64, 256, 512, 1024, 2048

Table 2. Network architecture overview of tiny bird object detection algorithm for infrared images.

Network Stage	Module	Feature Resolution	Channel Number
Multiscale Feature Extraction	Backbone Branch	256×256, 128×128, 64×64, 32×32, 16×16, 8×8	64, 128, 256, 512, 512, 512
	Detail-preservation Branch	128×128, 64×64, 32×32	128, 256, 512
	Multiscale Feature Fusion Branch	256×256, 128×128, 64×64, 32×32, 16×16	64, 128, 256, 512, 512
Context Information Extraction	Local Context extraction	128×128, 64×64, 32×32, 16×16, 8×8	128, 256, 512, 512, 512
	Global Context extraction	128×128, 64×64, 32×32, 16×16, 8×8	384, 768, 1536, 1536, 1536
Multiscale Prediction Supervision	Feature Fusion	512×512	7
	Output Head	512×512	1

4.2. Evaluation metric

For bird detection experiments in visible images, this paper AP and optimal localization recall precision ($oLRP$) [46] to evaluate the algorithms. In order to better describe the effectiveness of the proposed algorithm for detecting tiny bird targets, the IoU thresholds of 0.5 and 0.75 are chosen to calculate AP_{50} and AP_{75}, respectively. Inspired by the definitions used in the tiny object detection in aerial images (AI-TOD) dataset [47], which, in images with a resolution of 800×800 pixels, are categorized based on their pixel sizes: very tiny (2–8 pixels), tiny (8–16 pixels), small (16–32 pixels), and medium (32–64 pixels). For each size category, AP is computed separately and labeled as AP_{vt}, AP_t, AP_s, AP_m, respectively.

For bird detection experiments in infrared images, this paper adopts accuracy (Acc), IoU, normalized intersection over union ($nIoU$), probability of detection (P_d) and false-alarm rate (F_a) to evaluate the detection performance of the methods on infrared images. The IoU and $nIoU$ can be calculated by

$$IoU = \frac{A_i}{A_u}, nIoU = \frac{1}{N}\sum_{i}^{N}\frac{TP[i]}{T[i] + P[i] - TP[i]}, \tag{4.1}$$

where A_i and A_u denote the areas of the intersection region and union region, respectively. N represents the number of samples. $TP[\cdot]$ denotes the number of true positive pixels. $T[\cdot]$ denotes the number of the ground truth, and $P[\cdot]$ denotes the number of predicted positive pixels.

4.3. Results of experiments in visible images

4.3.1. Results of comparative experiments

The visualization results of the comparative experiment are shown in Figure 11. Considering the performance differences of different methods, only some methods with good detection effect are selected for display. In the resulting images, green boxes, yellow boxes, and red boxes are used to mark false positive (FP), false negative (FN), and true positive (TP). Additionally, the TP areas are magnified in the corners of the result images to more clearly observe the confidence of the detection

boxes. The confidence levels of detections vary among the different methods. Notably, the object detection method of FasterHRNet+NwdRka demonstrates the highest confidence. However, the occurrence of false positives also indicates its limited ability to distinguish tiny targets accurately. Benefiting from unique network architecture design and guided by contextual information, the proposed algorithm achieves accurate detection of tiny and inconspicuous bird targets. Compared to birds on the ground, the detection performance of algorithms for flying birds is more demanding due to the variations in morphology caused by shooting angles and wing flapping. As illustrated in the fourth row of Figure 11, most methods have difficulty in capturing features when dealing with objects with significant morphological variants, leading to more instances of both missed detections and false alarms in the results. The FasterHRNet+NwdRka method, which performs well in detecting ground-based bird targets, experiences a significant decrease in performance when detecting flying bird targets, possibly due to its limited generalization ability.

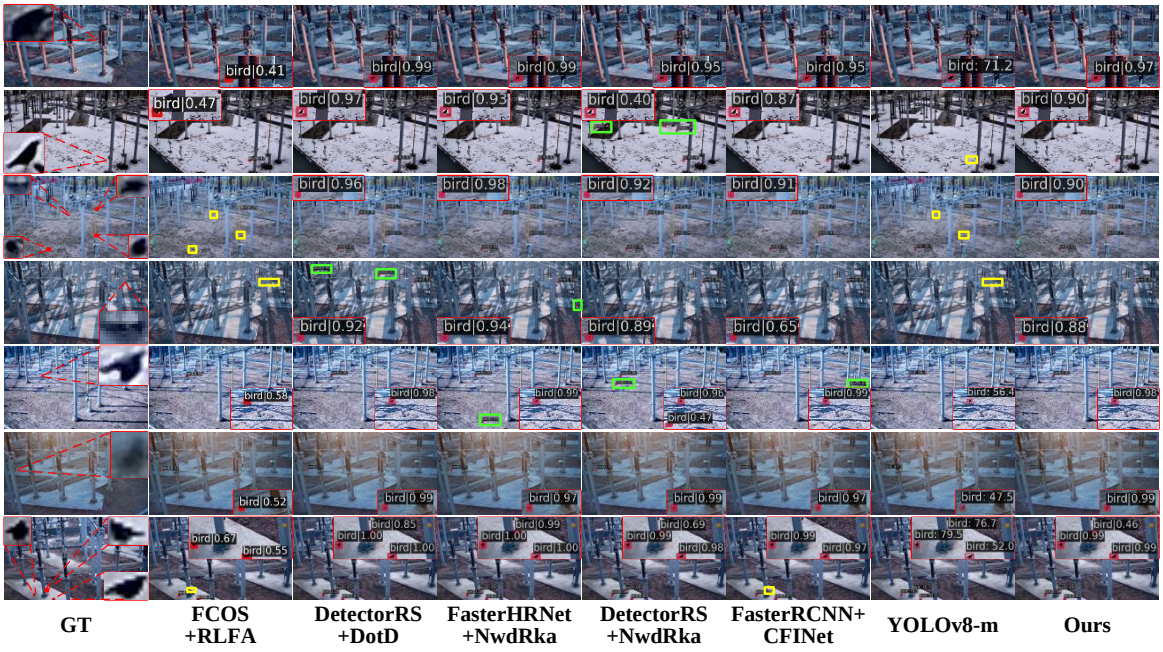


Figure 11. Comparison of bird targets detection effects in the VSTBD dataset.

The AP indicators calculated by the proposed method and all the comparison methods on VSTBD are shown in Table 4. Experimental results demonstrate that the overall AP of the proposed algorithm reaches 59.8%, representing a 2.5% improvement over the suboptimal FCOS+RLFA algorithm (AP=57.3%), with the AP_{75} metric increasing by 1.9%. Moreover, the proposed method achieves best performance across target scales of tiny (AP_t=51.4%), small (AP_s=62.5%), and medium (AP_m=69.7%). This superior performance can be attributed to the CIEF module, which effectively highlights tiny target regions in images, coupled with the HFE module that eliminates interference from non-avian targets. YOLOX-s, YOLOv8-s, and YOLOv8-m exhibited inferior performance, primarily attributed to the complexity and uniqueness of our experimental scenario as well as the extremely small size of the detected targets. These one-stage detectors struggle to capture sufficient discriminative features. Furthermore, one-stage detectors prioritize inference speed over the

depth of feature extraction, rendering them incapable of effectively modeling the contextual information of tiny bird targets. By contrast, our proposed method achieves outstanding performance, benefiting from two rounds of multiscale feature extraction on the input images and the integration of contextual information to enhance the features of tiny bird target regions.

Table 3. Main hyperparameters of the proposed algorithms for visible and infrared image detection.

Network	Module / Parameter	Configuration Details
Detection network for visible images	ResNet-50	Convolution Kernel: 1×1, Stride 1 Convolution Kernel: 3×3, Stride of Downsampling Block: 2 Convolution Kernel: 3×3, Stride of Normal Block: 1
	CIEF Module	Standard Convolution: 1×1 , Stride: 1 Dilated Convolution: 3×3, Dilation Rate: 3 , Stride: 1
	HFE Module	Convolution Kernel: 3×3, Stride: 1 Convolution Kernel: 5×5, Stride: 1 Convolution Kernel: 7×7, Stride: 1 Convolution Kernel: 9×9, Stride: 1
	FFPN Module	Dilated Convolution: 3×3, Dilation Rate: 3 , Stride: 1
	Spatial Attention	Convolution Kernel: 7×7
	Channel Attention	MLP Hidden Layer Dimension: 128
	NWD	Scaling Factor: 1, Matching Threshold: 0.5
	RKA	Positive Sample IoU Threshold: 0.5, Negative Sample IoU Threshold: 0.1, Sorting Top-k Value: 3
	Learning Rate	Initial: 0.01, Decay to 1% of Initial Value at the 9th and 11th Training Epochs
	Batch Size	4
	Weight Decay Coefficient	0.0001
	Momentum Coefficient	0.9
	Optimizer	Stochastic Gradient Descent
Detection network for infrared images	Backbone Branch	Depthwise Separable Convolution Kernel: 3×3, Stride: 1 Downsampling Stride: 2
	Fusion Block	Convolution Kernel: 3×3, Stride: 1
	Local Context Branch	Convolution Kernel: 1×1, Stride: 1, Convolution Kernel: 3×3, Stride: 1
	Window Attention	Window Size: 8, Number of Heads: 4
	Channel Attention	1D Convolution Kernel Size: 1×1
	Spatial Attention	Depthwise Convolution Kernel: 3×3
	Transformer	Number of Layers: 2, Feature Dimension per Attention Head: 256 Total Feature Dimension: 1024
	Learning Rate	Initial: 5×10^{-4}, Decay to 10% of Initial Value at the 200th and 300th Training Epochs
	Batch Size	32
	Optimizer	Adam

Since the network proposed in this paper requires iterative feature processing and multilevel feature enhancement, its parameter count is higher than that of other comparative methods. However, by adopting a relatively simple feature extraction backbone and prediction head structure, the computational complexity of the proposed algorithm does not increase substantially. As presented in Table 5, when deployed on the embedded industrial computer, the proposed algorithm achieves a processing speed of 12 frames per second (FPS), which can basically meet the requirements of practical applications.

Table 4. Quantitative experimental results of small bird detection on the VSTBD dataset.

Metric	AP	AP_{50}	AP_{75}	AP_{vt}	AP_t	AP_s	AP_m	$oLRP$	$oLRP_{IoU}$	$oLRP_{FP}$	$oLRP_{FN}$
FasterRCNN+RLFA	53.4	81.7	64.8	15.9	48.7	56.1	62.2	57.1	17.4	25.7	18.0
CascadeRCNN+RLFA	54.5	81.1	65.7	19.5	49.5	56.7	63.6	56.9	15.9	28.0	16.3
DetectorRS+RLFA	55.7	83.8	67.0	23.9	47.0	58.9	66.4	54.6	16.9	23.0	15.3
FCOS+RLFA	57.3	88.4	67.9	**26.2**	51.2	60.2	62.1	52.3	**16.4**	20.0	14.5
FasterRCNN+DotD	53.9	85.0	62.6	26.8	47.1	55.9	64.6	56.3	17.6	22.6	16.3
FasterHRNet+DotD	50.6	82.8	57.3	23.6	44.2	52.6	60.7	59.1	18.6	23.4	18.7
DetectorRS+DotD	53.7	86.0	61.3	22.5	46.9	56.2	64.7	56.4	18.0	20.1	17.7
FasterRCNN+NwdRka	52.0	84.5	59.1	22.4	45.3	54.3	62.7	57.8	18.5	24.8	14.8
CascadeRCNN+NwdRka	52.8	83.2	60.8	21.6	43.9	55.4	64.8	57.8	17.9	24.8	16.1
FasterHRNet+NwdRka	54.0	86.0	62.5	27.4	47.0	56.0	63.7	55.4	17.8	23.2	**12.6**
RetinaNet+NwdRka	52.5	89.3	56.0	21.8	43.4	55.5	63.2	54.2	18.9	15.2	15.1
DetectorRS+NwdRka	53.8	86.2	61.1	20.5	45.6	56.6	63.1	57.0	18.4	21.6	16.2
FasterRCNN+CFINet	54.6	86.8	61.7	23.3	46.3	57.4	65.0	54.0	17.5	16.4	17.7
YOLOX-s	42.3	82.2	38.5	14.0	39.5	44.8	50.8	64.4	22.6	22.1	20.3
YOLOv8-s	40.4	68.6	43.6	5.0	22.7	46.6	62.1	68.4	18.5	29.2	28.1
YOLOv8-m	43.4	71.9	47.6	6.9	27.4	48.8	65.2	66.3	18.1	28.7	32.7
Ours	**59.8**	**89.9**	**69.8**	25.4	**51.4**	**62.5**	**69.7**	**49.7**	16.6	**14.1**	14.1
Ours std	0.021	0.016	0.016	0.016	0.021	0.014	0.017	0.017	0.010	0.020	0.010

Note: All metrics are presented in percentage terms; bold numbers are the best, and underscored are second best. Ours is the mean of the metrics obtained from training our network three times. Ours std is the standard deviation of the corresponding metric.

Table 5. Analysis of network parameters and inference speed of tiny object detection methods in visible images on the embedded platform.

Method	FasterRCNN	CascadeRCNN	FasterHRNet	YOLOX-s	YOLOv8-m	Ours
Parameter (M)	41.6	69.0	46.9	8.9	25.9	124.5
FLOPs (G)	70.3	74.5	102.7	13.3	39.4	76.7
FPS	13	11	9	40	24	12

4.3.2. Results of ablation experiment

In order to verify the effectiveness of each module proposed in this paper in the detection network, an ablation experiment is carried out on the VSTBD dataset to analyze the performance of each module. The feature fusion module of the two encoders is replaced by a simple feature summation, and only the two-feature extraction and pyramid enhancement modules are retained to build the baseline model of the ablation experiment. The proposed CIEF, HFE, and FFPN are gradually added to the baseline model to verify the effectiveness of each proposed module. The ablation experimental results of the proposed method in VSTBD dataset is shown in Table 6.

Table 6. Ablation experiment results of the proposed visible image bird detection method on the VSTBD dataset.

Ablation	AP	AP_{50}	AP_{75}	AP_{vt}	AP_t	AP_s	AP_m	$oLRP$	$oLRP_{IoU}$	$oLRP_{FP}$	$oLRP_{FN}$
Strategy 1	52.6	83.3	60.7	23.5	44.6	55.2	64.6	58.0	18.1	25.0	15.6
Strategy 2	53.8	86.2	62.7	18.1	<u>47.1</u>	56.4	62.8	56.0	18.4	21.4	14.1
Strategy 3	54.8	88.3	61.3	20.6	46.7	57.9	65.2	54.6	18.6	18.6	<u>13.3</u>
Strategy 4	<u>58.5</u>	<u>88.4</u>	<u>69.0</u>	**26.3**	**51.4**	<u>61.2</u>	<u>68.0</u>	<u>50.8</u>	**16.5**	<u>17.5</u>	**12.9**
ours	**59.8**	**89.9**	**69.8**	<u>25.4</u>	**51.4**	**62.5**	**69.7**	**49.7**	<u>16.6</u>	**14.1**	14.1

Note: Strategy 1: Baseline; Strategy 2: Baseline + CIEF; Strategy 3: Baseline + CIEF + HFE; Strategy 4: Baseline + CIEF + FFPN; Ours: Baseline + CIEF + HFE + FFPN. All metrics are presented in percentage terms; bold numbers are the best, <u>underscored</u> second best.

The proposed CIEF can effectively capture the difference between pixels and surrounding neighborhoods and highlight local salient objects. Therefore, after using this module, more tiny target clues can be mined from the image. As shown in Table 6, in the detection results of Strategy 2, although there is a certain improvement compared with the baseline model overall, the index AP_{vt} is lower than the baseline model. This phenomenon may be attributed to the fact that the CIEF module focuses on local neighborhood features, while very tiny targets themselves contain extremely limited pixel information, making the module highly susceptible to interference from background information in their neighborhoods. Consequently, the index AP_{vt} is lower than that of the baseline. Strategy 3 can effectively improve network detection performance compared to using only the CIEF module. However, judging from the overall AP and AP_t metrics, the CIEF module proposed in this paper enables the network to pay more attention to tiny target regions. At the same time, from the result indicators of Strategy 4, it can be seen that the performance indicators are not much different when the HFE module is included or not included. The HFE model only improves the detection performance of the network for bird targets with a small amount of calculation.

4.4. Results of experiments in infrared images

4.4.1. Results of comparative experiments

The qualitative analysis of the comparative experimental results of the tiny birds detection algorithm in the infrared image on IRTBD dataset is shown in the Figure 12. Green boxes, red boxes, and yellow boxes are used to mark TP, FP and FN. It can be seen that the algorithm proposed in this paper

can detect tiny bird targets in infrared images well. Other algorithms have false detection or missed detection. The picture contains a complex background, bird flight, and bad weather. Benefiting from the multiscale feature extraction and context information extraction module proposed in this paper, the features of small bird targets in the image can be well extracted, a the influence of complex background and bad weather on the detection accuracy can be reduced.

Based on the qualitative analysis of bird detection performance across different scenarios, this paper further evaluates the effectiveness of each algorithm using various objective quantitative metrics. As shown in Table 7, compared to other algorithms, the proposed algorithm achieves more comprehensive feature representation through dense connections of multilevel features, thereby improving its performance to a certain extent. The algorithm proposed in this paper achieves the optimal performance across all five metrics, with the metrics $PixAcc$, IoU, and $nIoU$ showing an average improvement of approximately 2.5% compared to the second-best algorithm. Furthermore, the algorithm attains the lowest F_a and the highest P_d, indicating that the proposed tiny bird detection algorithm in infrared images outperforms other advanced algorithms in distinguishing a tiny bird from heated electrical equipment in the complex environment of the transformer substation. Furthermore, the proposed algorithm shows a significant improvement in the critical IoU metric, demonstrating that it achieves more accurate localization of tiny targets in tiny object detection tasks.

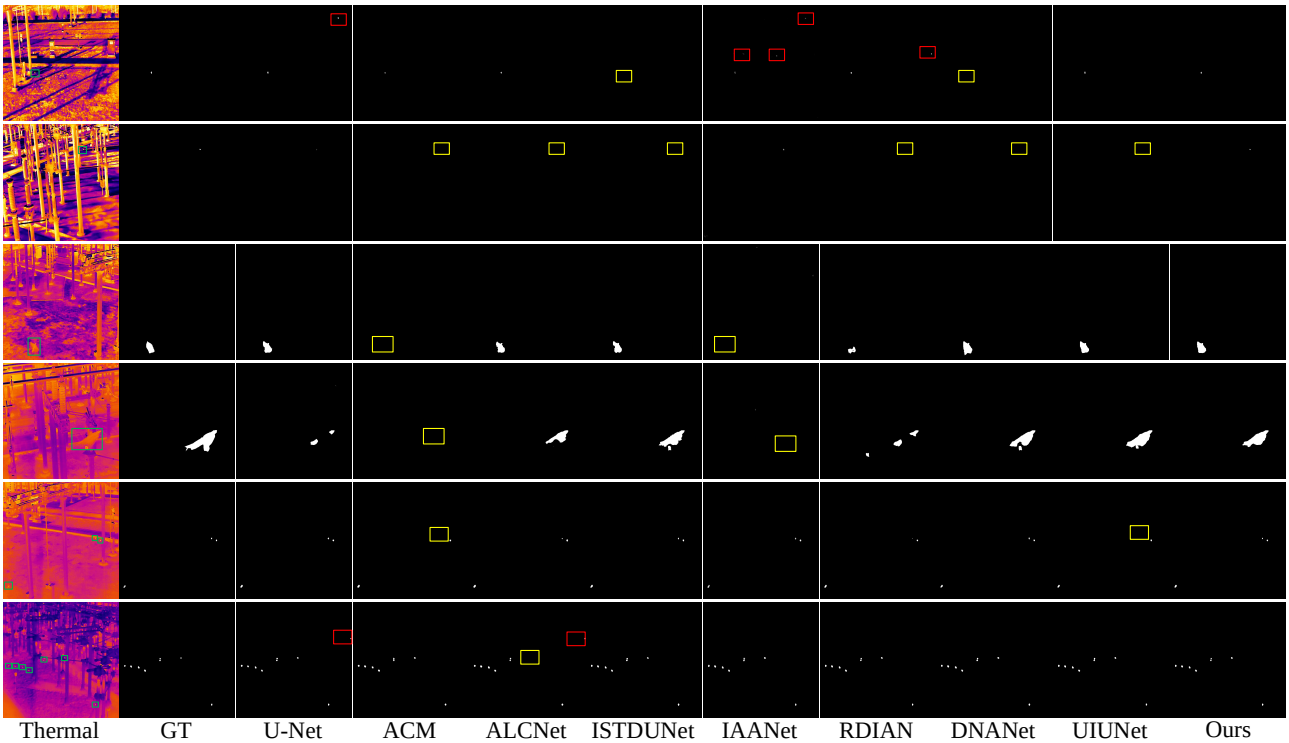


Figure 12. Detection results for bird targets in the IRTBD dataset.

Finally, a comparative analysis of the complexity and computational cost of infrared tiny object detection networks is conducted as presented in Table 8. DNANet adopts a densely connected U-Net architecture to achieve infrared tiny object detection, and it maintains a low parameter count and computational cost by reducing feature dimensions. Compared with other networks, the proposed

network exhibits a relatively low overall parameter count due to its simpler architectural design; however, considering the multiscale feature enhancement and simultaneous prediction mechanisms, its computational cost increases to a certain extent. When deployed on an embedded industrial computer, the proposed network achieves a processing rate of approximately 20 FPS, which can fully meet the requirements of practical applications.

Table 7. Quantitative experimental results of tiny bird detection on the IRTBD dataset.

Method	U-Net	ACM	ALCNet	ISTDUNet	IAANet	RDIAN	DNANet	UIUNet	Ours(Mean+std)
PixAcc	75.75	67.55	73.93	78.84	74.44	74.68	78.40	<u>79.33</u>	**81.24±0.028**
IoU	57.90	52.47	55.07	61.24	56.96	60.04	61.86	<u>63.18</u>	**65.82±0.021**
nIoU	57.55	51.95	51.95	59.60	59.13	60.10	60.56	<u>61.85</u>	**64.85±0.031**
P_d	84.46	74.83	78.56	84.00	<u>87.26</u>	84.62	84.71	84.06	**89.03±0.025**
$F_a(\times 10^3)$	7.73	<u>5.77</u>	8.23	6.74	7.82	6.16	6.79	6.56	**5.63±0.022**

Note: All metrics are presented in percentage terms; bold numbers are the best, <u>underscored</u> second best. Ours(Mean+std) are the results of the metrics obtained from training our network three times.

Table 8. Analysis of network parameters and inference speed of tiny object detection methods in infrared images on the embedded platform.

Method	U-Net	ACM	ALCNet	ISTDUNet	DNANet	UIUNet	Ours
Parameter (M)	34.53	39.78	42.70	44.47	4.70	50.54	44.47
FLOPs (G)	261.8	1.6	1.5	30.8	57.0	217.7	108.9
FPS	14	48	42	35	32	15	20

4.4.2. Results of ablation experiment

To verify the effectiveness of each module in the proposed infrared image bird object detection network, we constructed different models by removing distinct modules from the network on the IRTBD dataset. The experimental results are presented in Table 9. Compared with the baseline model, the newly added multiscale feature fusion branch fully trains multilevel features through cross-layer connections, leading to an approximate 3% improvement in the IoU metric. Meanwhile, the detail-preservation branch can transmit more detailed information to high-level features, thereby maintaining the effectiveness of synchronous prediction of multilevel features. When this branch is used independently, all metrics are improved to a certain extent. The simultaneous deployment of the backbone network, detail-preservation branch, and multiscale feature fusion branch can enhance the multiscale feature representation capability of network and significantly boost its overall performance. Furthermore, the ablation experiments integrating the contextual information extraction module and multiscale prediction supervision module demonstrate the validity of the module design through the significant improvements in $PixAcc$, IoU, and $nIoU$ metrics.

Table 9. Ablation experiment results of the proposed infrared tiny bird detection method.

Ablation procedure	$PixAcc$	IoU	$nIoU$	P_d	F_a
Extraction baseline (same decoder)	78.05	61.39	60.28	86.17	6.74
+Branch1	79.86	62.61	61.09	87.41	6.50
+Branch2	<u>80.18</u>	<u>64.31</u>	<u>62.91</u>	<u>88.19</u>	<u>6.12</u>
+Branch1 + Branch2	**81.24**	**65.82**	**64.85**	**89.03**	**5.63**
Decoder baseline (same encoder)	61.77	39.66	37.49	65.38	11.28
+Module1	67.51	42.90	37.79	65.54	10.50
+Module2	<u>77.62</u>	<u>60.15</u>	<u>60.25</u>	<u>86.95</u>	<u>7.69</u>
+Module1 + Module2	**81.24**	**65.82**	**64.85**	**89.03**	**5.63**

Note: Branch1 and Branch2 represent the detail-preservation branch and the multiscale feature fusion branch, respectively. Module1 and Module2 denote the contextual information extraction and multiscale prediction supervision, respectively. All metrics are presented in percentage terms; bold numbers are the best, <u>underscored</u> second best.

4.5. Results of bird target location

The infrared and visible cameras are calibrated in this section. The resolution of the infrared image is manually adjusted to 1920 × 1080 pixels, which is consistent with the visible image. A heated perforated calibration board was employed to perform cocalibration of the infrared and visible-light cameras, and the intrinsic and extrinsic parameters of the heterogeneous binocular camera were calculated. The calibrated binocular camera parameters are presented in Table 10. Based on the calibration parameters of the left and right heterogeneous cameras, the Bouguet stereo correction algorithm is used to correct heterogeneous images, and all calibration board images acquired during the calibration process were visualized for verification. To quantitatively evaluate the calibration accuracy of the binocular camera system, the mean reprojection error was adopted to assess the calibration performance of the binocular images, as illustrated in Figure 13.

Table 10. Basic parameters of heterogeneous stereo cameras.

Camera Parameters	Left Camera (Visible Camera)	Right Camera (Infrared Camera)
Intrinsic Matrix	$\begin{bmatrix} 2551.29 & -14.77 & 951.14 \\ 0 & 2568.36 & 335.55 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 4623.85 & -14.46 & 906.18 \\ 0 & 3264.01 & 370.78 \\ 0 & 0 & 1 \end{bmatrix}$
Distortion Coefficients	$\begin{bmatrix} -0.18 & -14.97 & -0.002 & -0.01 & 269.47 \end{bmatrix}$	$\begin{bmatrix} -0.40 & 8.85 & -0.009 & -0.01 & -111.97 \end{bmatrix}$
Calibrated Projection Matrix	$\begin{bmatrix} 2916.19 & 0 & -46.05 & 0 \\ 0 & 2916.19 & 429.33 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 2916.19 & 0 & -46.05 & -319444.58 \\ 0 & 2916.19 & 429.33 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$
Binocular Rotation Matrix	$\begin{bmatrix} 0.997371297873745 & -0.00162502879483336 & 0.0724420696767953 \\ -0.00146692338845179 & 0.999090785857233 & 0.0426080948993218 \\ -0.0724454437036182 & -0.0426023578759810 & 0.996462089991382 \end{bmatrix}$	
Binocular Translation Vector	$\begin{bmatrix} -105.445042960864 & -0.0577045977877141 & -29.6776795792478 \end{bmatrix}^T$	

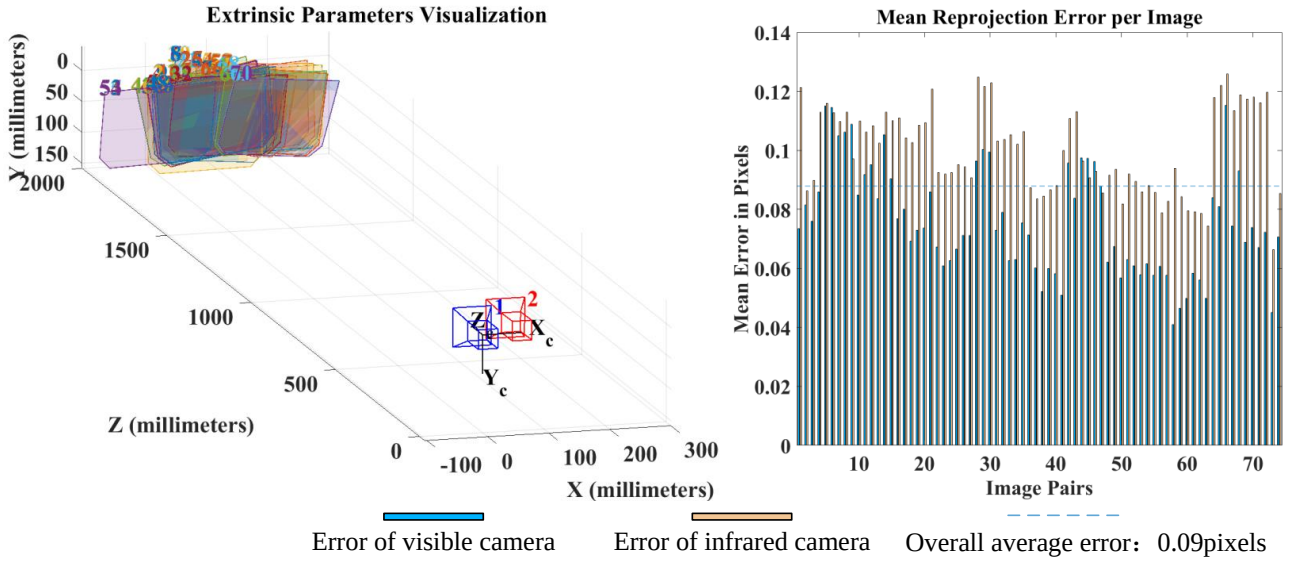


Figure 13. Calibration process and reprojection error of heterogeneous stereo cameras.

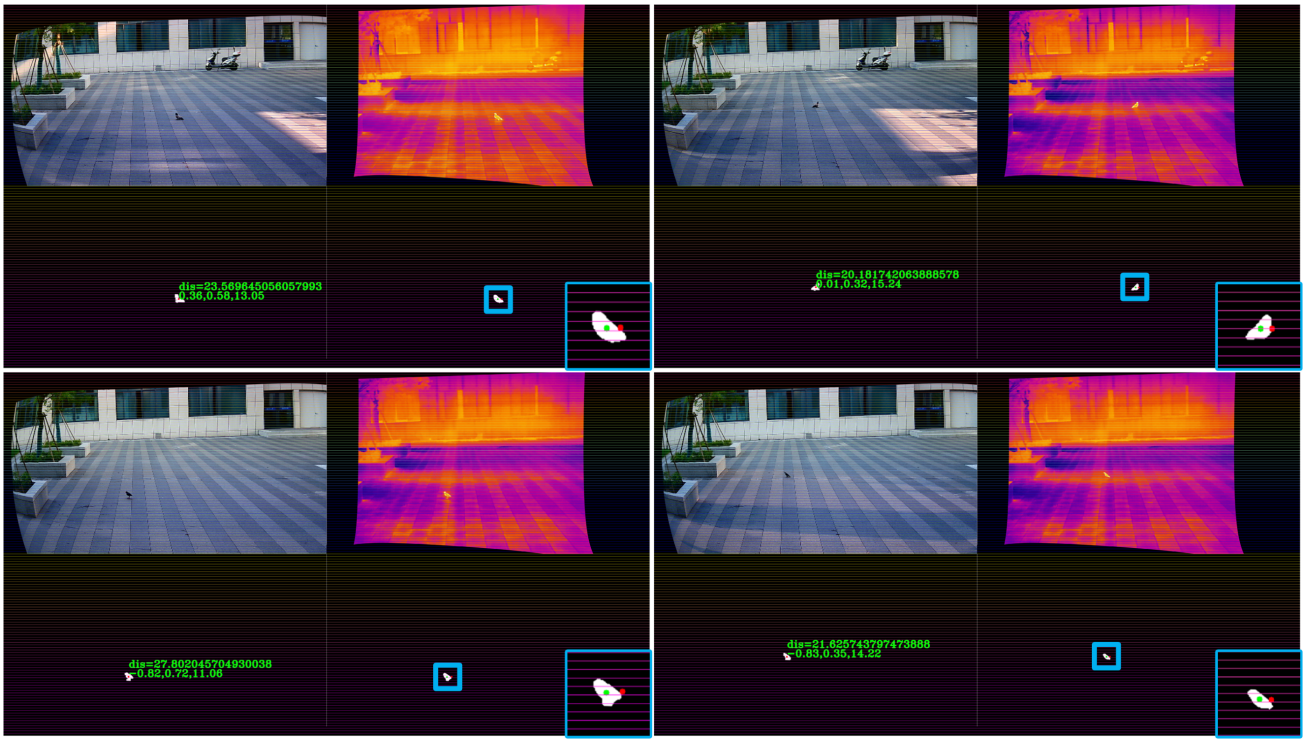


Figure 14. Bird detection and localization results in test scenarios. Green dots and red dots in the figure represent the pixel centers of bird targets in infrared images and visible images, respectively.

After camera calibration, the center pixel of the bird target is determined using the baseline constraint of the binocular camera. After calculating the bird's parallax, the distance between the bird

target and the device is measured, thus achieving spatial localization of the bird target. The experimental results of the bird target distance measurement are shown in Figure 14. The pixel centers of bird targets in infrared and visible images are marked with green and red dots respectively. The calculated object disparity and three-dimensional positioning results are marked in the visible image detection results. It can be clearly seen that for bird targets of different distances and shapes, the center of the bird target in the heterogeneous image has only a tiny longitudinal error. Therefore, the same object in the heterogeneous image is determined based on the baseline constraint and the disparity is calculated. To further verify the accuracy of distance measurement, heterogeneous images of bird targets were collected at different ranges, which were categorized into short-range (5–10 m), medium-range (10–15 m), and long-range (15–20 m). The depth values acquired by the Livox AVIA LiDAR were adopted as the ground truth to quantitatively evaluate the performance of bird target localization via heterogeneous binocular vision. The results of the quantitative metrics are presented in Table 11. As can be observed from the table, the localization errors of all targets are negligible except for bird 4. The complex morphology and large size of bird 4 led to a deviation in the pixel center localization. However, the resulting measurement error of 4.9% still satisfies the requirements of practical engineering applications.

Table 11. Measurement error analysis of heterogeneous bird target localization.

Experimental group	Distance category	Radar measurement true value (m)	Heterogeneous binocular measurement result (m)	Relative error
bird 1	short-range	6.82	6.89	1.03%
bird 2	short-range	7.35	7.42	0.95%
bird 3	short-range	8.56	8.71	1.75%
bird 4	short-range	9.75	10.23	4.90%
bird 5	medium-range	11.31	11.11	1.80%
bird 6	medium-range	12.94	13.06	0.90%
bird 7	medium-range	13.42	13.50	0.60%
bird 8	medium-range	14.18	13.96	1.55%
bird 9	long-range	15.25	15.24	0.20%
bird 10	long-range	16.95	17.12	1.00%
bird 11	long-range	18.32	18.05	1.47%
bird 12	long-range	19.56	20.03	2.40%
Std	-	-	-	1.168%

Note: Experimental results across three distance categories (4 groups per category), totaling 12 groups: short-range (5–10 m), medium-range (10–15 m), and long-range (15–20 m).

5. Conclusions

In this paper, a tiny bird detection and location system guided by heterogeneous binocular images was developed which consists of an embedded industrial computer, a heterogeneous binocular module, and a laser transmitter. The system integrated two algorithms proposed by this paper to enable tiny birds detection not only in the daytime but also under low light conditions. A two-stage contextual information enhancement network is proposed to detect birds at the scene with sufficient light, which

uses the context information in the primary coding feature to assist the coding process of the secondary feature and increase the difference between the bird target and the background area in the feature. A multiscale contextual feature enhancement network is proposed to detect birds at the low light scene. A multibranch feature coding network is constructed to extract differentiated multiscale information. Then, the multiscale feature context information is extracted and used for prediction, making full use of the information from features in different levels. The bird target segmentation is realized on the basis of accurate positioning of the bird target. Finally, the bird targets in different images are matched, and their spatial information is calculated based on the heterogeneous binocular module. Experiments on self-made datasets verify the effectiveness of the algorithms proposed in this paper and show that the developed system meets the robustness and accuracy requirements of birds detection and location in practical engineering.

Despite these achievements, this study has key limitations. First, real-time performance: While the 12–20 FPS on embedded computers meets basic needs, the two-stage and multibranch structures restrict ultra-high real-time applications. Second, hardware cost: High-resolution heterogeneous binocular modules limit large-scale deployment in cost-sensitive scenarios. Third, extreme weather robustness: Performance may degrade under heavy rain, fog, or strong sunlight due to environmental noise interference. In future work, we plan to explore algorithm lightweighting, low-cost sensor integration, optimized fusion, and the expansion of datasets to include extreme weather.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the Ningxia Natural Science Foundation (Grant No. 2024AAC03749) and the Science and Technology Project of Ultrahigh Voltage Company, State Grid Ningxia Electric Power Co., Ltd. (Contract No. SGNXCG00YCJS2400594).

Conflict of interest

The authors declare there are no conflicts of interest.

References

1. H. Zhou, H. Zhang, A. Wang, X. Fang, H. Xu, Y. Song, Application of intelligent bird repelling technology for power transmission and transformation equipment, in *2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, (2024), 289–293. https://doi.org/10.1109/ICPICS62053.2024.10796329
2. Q. Chen, J. Xie, Q. Yu, C. Liu, W. Ding, X. Li, et al., An experimental study of acoustic bird repellents for reducing bird encroachment in pear orchards, *Front. Plant Sci.*, **15** (2024), 1365275. https://doi.org/10.3389/fpls.2024.1365275

3. S. Shanmugapriya, S. Srikanth, S. Stanley, S. Surya, Intelligent IoT system for animal detection and crop defense using acoustic and visual repellents, in *2025 3rd International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA)*, (2025), 1–5. https://doi.org/10.1109/AIMLA63829.2025.11040691

4. Y. Chen, J. Chu, K. Hsieh, T. Lin, P. Chang, Y. Tsai, Automatic wild bird repellent system that is based on deep-learning-based wild bird detection and integrated with a laser rotation mechanism, *Sci. Rep.*, **14** (2024), 15924. https://doi.org/10.1038/s41598-024-66920-2

5. J. Wu, Z. Pan, B. Lei, Y. Hu, FSANet: Feature-and-spatial-aligned network for tiny object detection in remote sensing images, *IEEE Trans. Geosci. Remote Sens.*, **60** (2022), 1–17. https://doi.org/10.1109/TGRS.2022.3205052

6. D. Liu, J. Zhang, Y. Qi, Y. Wu, Y. Zhang, A tiny object detection method based on explicit semantic guidance for remote sensing images, *IEEE Geosci. Remote Sens. Lett.*, **21** (2024), 1–5. https://doi.org/10.1109/LGRS.2024.3374418

7. L. Xu, Y. Song, W. Zhang, Y. An, Y. Wang, H. Ning, An efficient foreign objects detection network for power substation, *Image Vision Comput.*, **109** (2021), 104159. https://doi.org/10.1016/j.imavis.2021.104159

8. J. Ou, J. Wang, J. Xue, J. Wang, X. Zhou, L. She, et al., Infrared image target detection of substation electrical equipment using an improved faster R-CNN, *IEEE Trans. Power Delivery*, **38** (2023), 387–396. https://doi.org/10.1109/TPWRD.2022.3191694

9. S. Qiao, L. Chen, A. Yuille, DetectoRS: Detecting objects with recursive feature pyramid and switchable atrous convolution, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 10208–10219. https://doi.org/10.1109/CVPR46437.2021.01008

10. M. Yue, L. Zhang, J. Huang, H. Zhang, Lightweight and efficient tiny-object detection based on improved yolov8n for UAV aerial images, *Drones*, **8** (2024), 276. https://doi.org/10.1109/aiac63745.2024.10899550

11. B. Mahaur, K. K. Mishra, A. Kumar, An improved lightweight small object detection framework applied to real-time autonomous driving, *Expert Syst. Appl.*, **234** (2023), 121036. https://doi.org/10.1016/j.eswa.2023.121036

12. Z. Zhao, J. Du, C. Li, X. Fang, Y. Xiao, J. Tang, Dense tiny object detection: A scene context guided approach and a unified benchmark, *IEEE Trans. Geosci. Remote Sens.*, **62** (2024), 1–13. https://doi.org/10.1109/TGRS.2024.3357706

13. J. Xiao, H. Guo, J. Zhou, T. Zhao, Q. Yu, Y. Chen, et al., Tiny object detection with context enhancement and feature purification, *Expert Syst. Appl.*, **211** (2023), 118665. https://doi.org/10.1016/j.eswa.2022.118665

14. S. Huang, C. Lin, X. Jiang, Z. Qu, BRSTD: Bio-inspired remote sensing tiny object detection, *IEEE Trans. Geosci. Remote Sens.*, **62** (2024), 1–15. https://doi.org/10.1109/TGRS.2024.3470900

15. H. Liu, Y. Tseng, K. Chang, P. Wang, H. Shuai, W. Cheng, A denoising FPN with transformer R-CNN for tiny object detection, *IEEE Trans. Geosci. Remote Sens.*, **62** (2024), 1–15. https://doi.org/10.1109/TGRS.2024.3396489

16. J. Leng, Y. Ren, W. Jiang, X. Sun, Y. Wang, Realize your surroundings: Exploiting context information for small object detection, *Neurocomputing*, **433** (2021), 287–299. https://doi.org/10.1016/j.neucom.2020.12.093

17. D. Ma, B. Liu, Q. Huang, Q. Zhang, MwdpNet: Towards improving the recognition accuracy of tiny targets in high-resolution remote sensing image, *Sci. Rep.*, **13** (2023), 13890. https://doi.org/10.1038/s41598-023-41021-8

18. C. Xu, J. Ding, J. Wang, W. Yang, H. Yu, L. Yu, et al., Dynamic coarse-to-fine learning for oriented tiny object detection, in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2023), 7318–7328. https://doi.org/10.1109/CVPR52729.2023.00707

19. D. Liu, J. Zhang, Y. Qi, Y. Wu, Y. Zhang, A tiny object detection method based on explicit semantic guidance for remote sensing images, *IEEE Trans. Geosci. Remote Sens.*, **21** (2024), 1–5. https://doi.org/10.1109/LGRS.2024.3374418

20. C. W. Corsel, M. van Lier, L. Kampmeijer, N. Boehrer, E. M. Bakker, Exploiting temporal context for tiny object detection, in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, (2023), 1–11. https://doi.org/10.1109/WACVW58289.2023.00013

21. B. Li, C. Xiao, L. Wang, Y. Wang, Z. Lin, M. Li, et al., Dense nested attention network for infrared small target detection, *IEEE Trans. Image Process.*, **32** (2022), 1745–1758. https://doi.org/10.1109/TIP.2022.3199107

22. X. Wu, D. Hong, J. Chanussot, UIU-Net: U-Net in U-Net for infrared small object detection, *IEEE Trans. Image Process.*, **32** (2023), 364–376. https://doi.org/10.1109/TIP.2022.3228497

23. T. Zhang, L. Li, S. Cao, T. Pu, Z. Peng, Attention-guided pyramid context networks for detecting infrared small target under complex background, *IEEE Trans. Aerosp. Electron. Syst.*, **59** (2023), 4250–4261. https://doi.org/10.1109/TAES.2023.3238703

24. L. Huang, S. Dai, T. Huang, X. Huang, H. Wang, Infrared small target segmentation with multiscale feature representation, *Infrared Phys. Technol.*, **116** (2021), 103755. https://doi.org/10.1016/j.infrared.2021.103755

25. X. Xiong, K. Wang, J. Chen, T. Li, B. Lu, F. Ren, A calibration system of intelligent driving vehicle mounted scene projection camera based on zhang zhengyou calibration method, in *2022 34th Chinese Control and Decision Conference (CCDC)*, (2022), 747–750. https://doi.org/10.1109/CCDC55256.2022.10034031

26. C. Xu, J. Wang, W. Yang, H. Yu, L. Yu, G. Xia, Detecting tiny objects in aerial images: A normalized Wasserstein distance and a new benchmark, *ISPRS J. Photogramm. Remote Sens.*, **190** (2022), 79–93. https://doi.org/10.1016/j.isprsjprs.2022.06.002

27. J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, *IEEE Trans. Pattern Anal. Mach. Intell.*, **42** (2020), 2011–2023. https://doi.org/10.1109/TPAMI.2019.2913372

28. F. Chollet, Xception: Deep learning with depthwise separable convolutions, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 1800–1807. https://doi.org/10.1109/CVPR.2017.195

29. X. Chen, Y. Zhang, Y. Dong, B. Du, Spatial-spectral contrastive self-supervised learning with dual path networks for hyperspectral target detection, *IEEE Trans. Geosci. Remote Sens.*, **62** (2024), 1–12. https://doi.org/10.1109/TGRS.2024.3390946

30. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, et al., Swin transformer: Hierarchical vision transformer using shifted windows, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986

31. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Trans. Pattern Anal. Mach. Intell.*, **39** (2017), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

32. Z. Tian, C. Shen, H. Chen, T. He, FCOS: A simple and strong anchor-free object detector, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2022), 1922–1933. https://doi.org/10.1109/TPAMI.2020.3032166

33. Z. Cai, N. Vasconcelos, Cascade R-CNN: Delving into high quality object detection, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 6154–6162. https://doi.org/10.1109/CVPR.2018.00644

34. K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 5686–5696. https://doi.org/10.1109/CVPR.2019.00584

35. T. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, *IEEE Trans. Pattern Anal. Mach. Intell.*, **42** (2020), 318–327. https://doi.org/10.1109/TPAMI.2018.2858826

36. J. Terven, D. Córdova-Esparza, J. Romero-González, A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS, *Mach. Learn. Knowl. Extr.*, **5** (2023), 1680–1716. https://doi.org/10.3390/make5040083

37. Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, YOLOX: Exceeding YOLO series in 2021, preprint, arXiv:2107.08430.

38. C. Xu, J. Wang, W. Yang, H. Yu, L. Yu, G. Xia, RFLA: Gaussian receptive field based label assignment for tiny object detection, in *Computer Vision-ECCV 2022* , (2022), 526–543. https://doi.org/10.1007/978-3-031-20077-9_31

39. C. Xu, J. Wang, W. Yang, L. Yu, Dot distance for tiny object detection in aerial images, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 1192–1201. https://doi.org/10.1109/CVPRW53098.2021.00130

40. X. Yuan, G. Cheng, K. Yan, Q. Zeng, J. Han, Small object detection via coarse-to-fine proposal generation and imitation learning, in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2023), 6294–6304. https://doi.org/10.1109/ICCV51070.2023.00581

41. Y. Dai, Y. Wu, F. Zhou, K. Barnard, Asymmetric contextual modulation for infrared small target detection, in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, (2021), 949–958. https://doi.org/10.1109/WACV48630.2021.00099

42. Y. Dai, Y. Wu, F. Zhou, K. Barnard, Attentional local contrast networks for infrared small target detection, *IEEE Trans. Geosci. Remote Sens.*, **59** (2021), 9813–9824. https://doi.org/10.1109/TGRS.2020.3044958

43. Q. Hou, L. Zhang, F. Tan, Y. Xi, H. Zheng, N. Li, ISTDU-Net: Infrared small-target detection U-Net, *IEEE Geosci. Remote Sens. Lett.*, **19** (2022), 1–5. https://doi.org/10.1109/LGRS.2022.3141584

44. K. Wang, S. Du, C. Liu, Z. Cao, Interior attention-aware network for infrared small target detection, *IEEE Trans. Geosci. Remote Sens.*, **60** (2022), 1–13. https://doi.org/10.1109/TGRS.2022.3163410

45. H. Sun, J. Bai, F. Yang, X. Bai, Receptive-field and direction induced attention network for infrared dim small target detection with a large-scale dataset irdst, *IEEE Trans. Geosci. Remote Sens.*, **61** (2023), 1–13. https://doi.org/10.1109/TGRS.2023.3235150

46. K. Oksuz, B. C. Cam, E. Akbas, S. Kalkan, Localization recall precision (LRP): A new performance metric for object detection, in *Computer Vision-ECCV 2018*, **11211** (2018), 521–537. https://doi.org/10.1007/978-3-030-01234-2_31

47. J. Wang, W. Yang, H. Guo, R. Zhang, G. Xia, Tiny object detection in aerial images, in *2020 25th International Conference on Pattern Recognition (ICPR)*, (2021), 3791–3798. https://doi.org/10.1109/ICPR48806.2021.9413340

AIMS Press