remote sensing

MDPI

Article

AIMED-Net: An Enhancing Infrared Small Target Detection Net in UAVs with Multi-Layer Feature Enhancement for Edge Computing

Lehao Pan, Tong Liu *, Jianghua Cheng, Bang Cheng and Yahui Cai

College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China; lehaopan22@nudt.edu.cn (L.P.); jianghua_cheng@nudt.edu.cn (J.C.); chengbang@nudt.edu.cn (B.C.); caiyh95@nudt.edu.cn (Y.C.)
* Correspondence: liutong1129@nudt.edu.cn

Abstract: In the context of small unmanned aerial vehicles (UAVs), infrared imaging faces challenges such as low quality, difficulty in detecting small targets, high false alarm rates, and computational resource constraints. To address these issues, we introduce AIMED-Net, an enhancing infrared small target detection net in UAVs with multi-layer feature enhancement for edge computing. Initially, the network encompasses a multi-layer feature enhancement architecture for infrared small targets, including a generative adversarial-based shallow-feature enhancement network and a detection-oriented deep-feature enhancement network. Specifically, an infrared image-feature enhancement method is proposed for the shallow-feature enhancement network, employing multi-scale enhancement to bolster target detection performance. Furthermore, within the YOLOv7 framework, we have developed an improved object detection network integrating multiple feature enhancement techniques, optimized for infrared targets and edge computing conditions. This design not only reduces the model's complexity but also enhances the network's robustness and accuracy in identifying small targets. Experimental results obtained from the HIT-UAV public dataset indicate that, compared to YOLOv7s, our method achieves a 2.5% increase in F1 score, a 6.1% rise in AP for detecting OtherVehicle targets, and a 2.6% improvement in mAP across all categories, alongside a 15.2% reduction in inference time on edge devices. Compared to existing state-of-the-art approaches, our method strikes a balance between detection efficiency and accuracy, presenting a practical solution for deployment in aerial edge computing scenarios.

Keywords: edge computing; deep learning; infrared image enhancement; infrared small target detection

check for updates

Citation: Pan, L.; Liu, T.; Cheng, J.; Cheng, B.; Cai, Y. AIMED-Net: An Enhancing Infrared Small Target Detection Net in UAVs with Multi-Layer Feature Enhancement for Edge Computing. *Remote Sens.* **2024**, *16*, 1776. https://doi.org/10.3390/rs16101776

Academic Editor: Susana Lagüela López

Received: 11 March 2024
Revised: 1 May 2024
Accepted: 15 May 2024
Published: 17 May 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Infrared imaging technology, known for its capabilities such as all-weather functionality and excellent concealment, has been widely applied in fields including military reconnaissance, environmental monitoring, and security surveillance. In the context of aerial applications, infrared detection is advantageous due to its low sensitivity to lighting conditions. This feature enables consistent monitoring and target tracking during nighttime or adverse weather conditions, essential for successful reconnaissance, patrol operations, and human activity surveillance.

Unmanned Aerial Vehicle (UAV) edge computing involves the execution of computational tasks and data storage directly at the UAV nodes, thereby minimizing latency, reducing bandwidth consumption, and enhancing response times. This approach, independent of cloud computing or central data processing facilities, enables real-time data analysis, autonomous decision making based on processed information, and more efficient task execution. It is crucial for scenarios requiring rapid processing and immediate action, such as environmental monitoring, disaster response, and battlefield reconnaissance.

Deploying uncooled infrared sensors on small unmanned aerial vehicles (UAVs) presents specific limitations. These sensors, limited by their imaging devices and design, often produce images with insufficient brightness, lost details, and reduced contrast under complex weather conditions [1], affecting the performance of infrared target detection and identification tasks. Additionally, detecting and identifying small, low-contrast targets with weak emission signatures and suboptimal signal-to-noise ratios in intricate scenes is a significant challenge. The high latency of cloud-based processing and the substantial computational requirements for real-time local processing also pose major challenges for UAV edge devices with limited resources. Consequently, enhancing the accuracy of infrared target detection for UAVs has become a crucial research area, especially for accurately detecting and identifying small infrared targets in complex aerial environments, which has substantial practical significance.

According to the Microsoft Common Objects In Context (MS COCO) [2], targets smaller than 32×32 pixels are categorized as small. These targets' detection in infrared imagery is notably challenged by noisy and complex backgrounds. Deep learning-based object detection algorithms outperform traditional methods by obviating the need for manual feature design. These algorithms fall into two categories: two-stage, region proposal-based methods like R-CNN [3] and its variants [4,5], which offer high accuracy but struggle with speed and efficiency; and single-stage methods such as SSD [6] and YOLO [7,8], optimized for speed and efficiency. Liu et al. [9] confirmed the effectiveness of Faster R-CNN in remote sensing target detection on the NWPU VHR dataset. Wu et al. [10] investigated the detection of military and civilian vehicles on UAV edge devices using optimized SSD and YOLOv3 structures.

Advancements in deep learning have significantly propelled its application in detecting small infrared targets. Liu et al. [11] were pioneers, creating a dataset to train predictive models by incorporating random target points into backgrounds with adjustable signal-to-noise ratios. McIntosh et al. [12] extended the utility of deep neural networks, including Faster R-CNN [5] and YOLOv3 [8], enhancing small infrared target detection. More recently, Park et al. [13] developed a CNN-based, pixel-level classifier surpassing traditional methods in nighttime infrared human detection. Hou et al. [14] introduced RISTDnet, a robust network for small infrared target detection, merging handcrafted features with CNNs. Zhao et al. [15] offered a novel approach using Generative Adversarial Networks specifically designed for small infrared targets. Zhang [16] proposed the Self-Regularized Weighted Sparsity (SRWS) model, framing detection as an optimization problem. Dai et al. [17] engineered an Asymmetric Context Module (ACM) to refine downsampling, attention mechanisms, and feature fusion. Leveraging ACM, they launched ALCNet [18], a network that amalgamates discriminative networks with model-driven methods, achieving unprecedented detection accuracy.

Furthermore, enhancing the quality of infrared images is a widely adopted strategy to improve the precision of infrared target detection. By enhancing the visibility of images of inferior quality, thereby producing images that are both clear and rich in detail aligned with human visual perception, this approach lays the groundwork for advanced visual tasks. Concurrently, within the context of applications requiring drone operators to possess a comprehensive understanding of their immediate surroundings, image enhancement emerges as a critical component for augmenting flight operations. Such enhancements, by delivering clearer and more intricate visual data, markedly elevate the efficacy and safety of drone flights. Li et al. [19] have introduced a deep learning strategy that integrates super-resolution techniques with YOLO to detect small targets in infrared remote sensing images. However, the super-resolution process might introduce artifacts, escalating computational and storage demands, and this could consequently prolong processing times. Such factors might impede its applicability in scenarios demanding a real-time response or where resources are limited. Liu et al. [20] proposed a method for detecting small infrared targets based on Transformer models, while Zhang et al. [21] developed an attention-guided contextual block to probe local semantic links and global context attention

mechanisms. This block is designed to identify pixel correlations within and across patches at specific scales, effectively detecting small infrared targets amidst complex backgrounds. Nonetheless, deploying and inferring models based on Transformer architectures and attention mechanisms on FPGA edge devices' DPU units can be challenging due to structural complexities, limiting their practical utility. Additionally, Ye et al. [22] proposed a two-stage deep learning model, Grid R-CNN, for detecting non-prominent targets in infrared images, showcasing superior performance. However, the complexity of the model's structure, the significant computational load, and redundancy in its operations may restrict its real-time applicability on edge devices.

Generally, neural networks can integrate image enhancement and target detection tasks through two approaches. The first approach trains an image enhancement network and uses the enhanced images to train a detection network. However, studies have shown that enhanced images do not always improve performance in high-level visual tasks, such as object detection [23], due to the divergent objectives of the two tasks. The second approach is a more integrated solution, combining both tasks into a single end-to-end network, such as the Hybrid-DetectionGAN. This model cascades a perceptual enhancement model with an SSD detector for joint training, generating images that facilitate detection in underwater scenarios. Lastly, Liu et al. [24] proposed a parallel connection of the enhancement and detection networks to enhance object detection performance in visually degraded environments.

Recent advancements have been achieved in the realms of infrared small target detection and image enhancement, marking initial progress in these areas. Nonetheless, the effective amalgamation of image enhancement processes with target detection tasks presents significant hurdles. Further complexities arise when attempting to fulfill the stringent requirements for high real-time performance and adherence to resource limitations within edge computing contexts.

The effectiveness of general algorithms in UAV-based infrared target detection is often compromised by the prevalence of small, indistinct targets in aerial imagery, leading to less than ideal detection performance. To address the limitations of current aerial infrared detection systems—including challenges in imaging quality, small target detection and identification, and processing efficiency—we propose AIMED-Net: An Aerial-based Integrated Network for Multi-layer Infrared Feature Enhancement and Small Target Detection, tailored for edge computing environments. We have deployed and tested this network on FPGA edge devices, demonstrating the suitability of our algorithm for practical application.

Our network architecture, specifically designed for infrared small target detection, incorporates a novel dual-layered feature enhancement approach that synergistically combines sequential and parallel enhancements to enhance both visual quality and depth-feature representation. This innovative configuration not only amplifies the visual and depth attributes but also seamlessly integrates image quality enhancement with target detection, achieving an end-to-end enhancement and detection mechanism that aligns with the stringent demands of edge computing environments. The architecture comprises both shallow and deep networks for comprehensive feature extraction, complemented by a state-of-the-art target detection network. This dual approach of sequential and parallel enhancements ensures robust feature enhancement and efficient detection tailored to the unique requirements of infrared imaging and edge computation.

While Liu et al. [25] adopted an enhancement-detection framework, our method introduces significant innovations. Specifically, it is optimized for identifying vehicles and humans in aerial infrared imagery and is tailored for UAV edge computing, unlike Liu et al.'s general focus on small airborne targets. Importantly, we pivot away from model-driven signal processing, which limits deployment and hardware acceleration on edge devices. Our approach also surpasses existing cascade methods, such as DSNet [26] and Hybrid-DetectionGAN [27], which struggle with infrared detection and offer limited applicability beyond their intended contexts. Unlike these methods, which primarily enhance detection performance without considering visual quality, our system enhances both

the visual quality for human operators and the detection accuracy, displaying improved imagery and detection outcomes simultaneously. This dual enhancement strategy ensures both operational efficiency and practical utility in real-world applications.

In summary, the principal contributions of our proposed method are as follows:

(1) We introduce AIMED-Net, a novel network optimized for UAV edge devices, integrating multi-layer infrared feature enhancement for small targets with an optimized YOLOv7 detection framework. This approach incorporates various feature enhancement strategies specifically tailored for infrared targets and edge computing environments.

(2) We propose an infrared image-feature enhancement technique that deliberately adds noise to specific regions of the original low-quality image. This degraded image serves as input, with the original image as ground truth (GT), for training the network. Through a generative adversarial network, the differences between the degraded input and the original image are compared, enabling the network to learn image restoration and target feature enhancement. Subsequently, the original images are iteratively fed into the network generator to produce higher-quality feature-enhanced images.

2. Materials and Methods

Inspired by the significant advancements in image enhancement and target detection, we propose a novel integrated network. This network is designed for multi-layer feature enhancement and the detection of small infrared targets, addressing the current limitations of airborne infrared detection systems, such as imaging quality, the detection and identification of weak and small targets, and processing efficiency. We have developed a lightweight framework with the YOLOv7 [28] object detection network's requirements in mind. This framework enhances the contrast and feature discernibility of infrared images without increasing computational complexity, providing high-quality input for detecting weak and small targets in airborne conditions. Additionally, by streamlining the data workflow with end-to-end processing, our approach minimizes data transmission and storage demands, optimizes resource allocation, and ensures efficient enhancement of image features, along with reliable detection of weak and small targets at the edge.

This section is organized into five parts: the overall architecture and principles of the model; the design of the loss function and training strategies; the high-altitude infrared dataset, HIT-UAV [29]; the deployment process in edge computing environments; and the evaluation metrics for validating the proposed methods in this study.

2.1. AIMED-Net

We introduce AIMED-Net, an advanced network designed for edge computing environments. AIMED-Net stands for Aerial-based Integrated Multi-layer Infrared Feature Enhancement and Small Target Detection Network, as depicted in Figure 1. This architecture is segmented into three primary components. The first is a shallow-feature enhancement network, utilizing adversarial generation to improve infrared imagery's visual characteristics. This is succeeded by a deep-feature enhancement network aimed at amplifying features of small targets, working in tandem with the target detection task. The final component is an optimized YOLOv7 target detection network, enhanced with the aforementioned techniques to suit edge computing scenarios. Both enhancement networks start from a pre-trained state, undergoing joint training with the target detection network to achieve a synergistic effect. This approach, supported by a combined loss function, ensures a robust linkage between feature enhancement and target detection. Consequently, it enables efficient and precise amplification of infrared image features and small target detection on airborne edge devices.

The AIMED-Net workflow is structured as follows: A shallow-feature enhancement network inputs low-resolution infrared images and outputs enhanced high-quality infrared images with finer details. These enhanced images are then simultaneously fed into the backbone of the object detection network and a deep-feature enhancement network for feature extraction and augmentation, respectively. Subsequently, the object detection back-

bone's extracted features are concatenated with the deep-feature enhancement network's enhanced features along the channel dimension. Finally, the combined features are input into the object detection network's prediction head, culminating in the display of the final detection results on the detail-enhanced high-quality images.

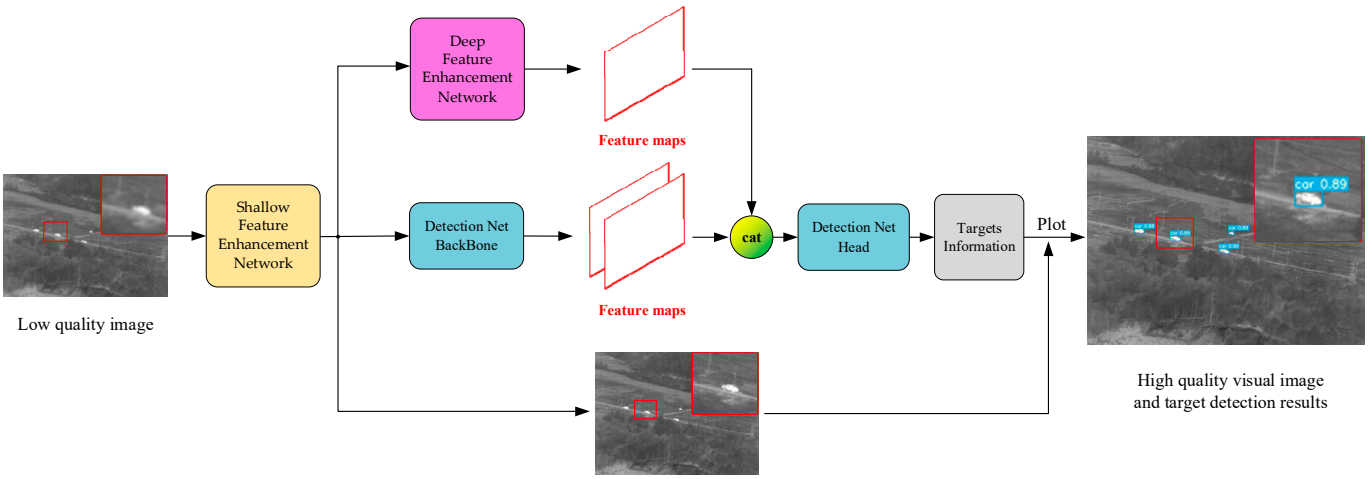


Figure 1. The architecture of AIMED-Net is presented, where the Detection Net Head adopts the YOLOv7 structure, while the detailed descriptions of other modules will be provided in subsequent sections.

Furthermore, given the limited computational resources and power supply characteristics of airborne edge devices, we have implemented a lightweight design within the YOLOv7 object detection network's detection component. This design is tailored to better suit such environments. Our approach enriches the images processed by the network with a wealth of semantic information through multiple stages of feature extraction, enhancement, and fusion within the network's backbone. Consequently, despite our network's streamlined architecture, its performance remains robust compared to architectures that feature deeper convolutional layers.

Our method effectively addresses the challenge of integrating image enhancement with object detection tasks. It enhances both the features of the objects and the visual outcomes.

The subsequent sections will detail the various components of AIMED-Net.

2.1.1. Generative Adversarial Network for Shallow-Feature Enhancement

Generative Adversarial Networks (GANs) [30] can significantly enhance infrared image quality through adversarial training between discriminator and generator networks. These networks, however, require a high number of training parameters, leading to increased resource consumption and computational complexity when deployed at the edge. To address this, we simplified the generator network's architecture by designing a shallow-feature extraction structure. This approach not only boosts feature extraction efficiency but also minimizes the network's layer count, effectively reducing the training parameters.

The architecture of our shallow-feature enhancement network is delineated in Figure 2, with Figure 2a detailing the generator network and Figure 2b outlining the discriminator network. The generator network deviates from the mainstream encoder-decoder architecture, adopting an isometric convolutional design aimed at achieving detail enhancement and denoising of infrared images, thereby minimizing information loss.

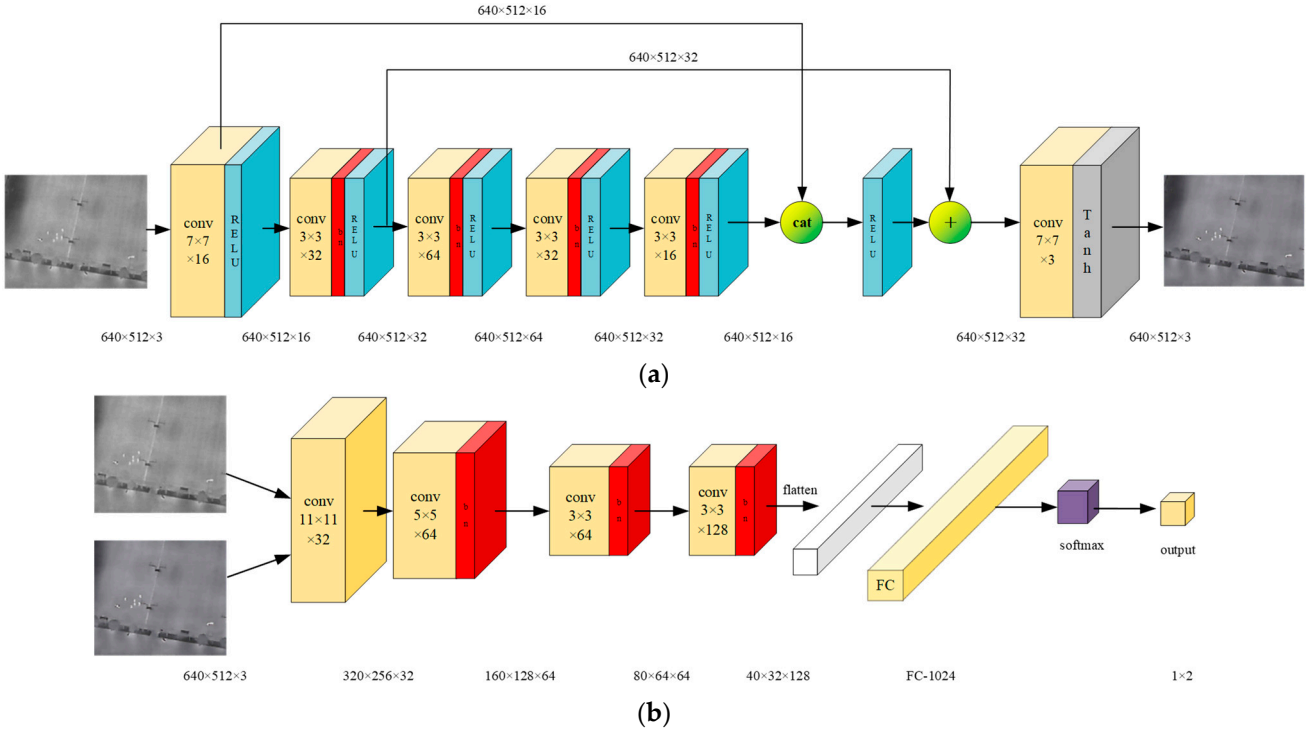


Figure 2. Architecture of a shallow-feature enhancement network Based on the Generative Adversarial Approach is presented. The BN (Batch Normalization) layers are represented by red segments, ReLU (Rectified Linear Unit) layers are depicted by blue segments, and Tanh (Hyperbolic Tangent) layers are indicated by gray segments. The figure includes annotations for parameters and feature map dimensions. (**a**) The structure of the shallow-feature enhancement generator network is illustrated; (**b**) the structure of the shallow-feature enhancement discriminator network is shown.

As illustrated in Figure 2a, we concatenate the output features from the first and fifth convolutional layers to preserve shallow image features. These features are then merged with those from the second layer via a weighted fusion method. Unlike the traditional approach of stacking convolutional layers, our architecture extracts multi-level image features with minimal computational overhead, thereby enhancing the efficiency of the convolutional layers in feature extraction. Differing from the conventional skip connections that employ weighted summation, our method minimizes the loss of shallow features. The generator network, consisting of six convolutional layers, extracts features across various scales. This design aids in the detailed comparison with high-quality target images in the discriminator network.

The discriminator network architecture employs a classic convolutional encoding structure. Figure 2b illustrates the discriminator network, comprising four convolutional layers and a fully connected neural network, with outcomes determined via a softmax layer. This network randomly processes either enhanced images from the generator or high-quality real images, classifying them as 'fake' or 'real', respectively. In the training phase, this serves to assist in the refinement of the generator network. During inference, the generator operates independently, ensuring the discriminator adds no extra computational complexity.

2.1.2. Deep-Feature Enhancement Network for Detection Tasks

Typically, image enhancement networks and target detection networks operate independently [31,32], resulting in the enhancement network's parameters not being optimized for improved detection performance. To address this issue, we designed a deep-feature enhancement network to augment target detection-relevant features.

As depicted in Figure 3, the deep-feature enhancement network consists of three convolutional layers, utilizing convolutional kernels of varied depths to extract image features across multiple scales. The output feature map of the final layer is then concatenated with that of a lightweight backbone along the channel dimension, enhancing feature representation. This combined feature map serves as the input for the target detection network's head section.

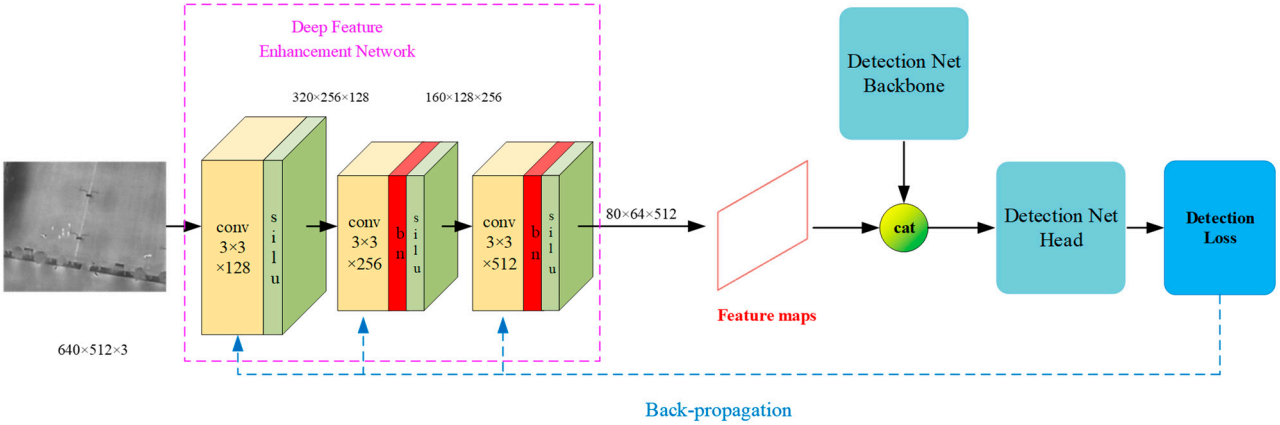


Figure 3. Structure of the deep-feature enhancement network is presented, where the Detection Loss denotes the loss function utilized for the target detection task, with comprehensive explanations provided in subsequent sections.

During the training phase, the deep-feature enhancement network is serially connected to the backbone and head sections, with their weights frozen to assist in optimizing the feature network's weights. This strategy ensures that the deep-feature enhancement network outputs image features during inference that are advantageous for the target detection network, effectively merging the tasks of image enhancement and target detection.

2.1.3. YOLOv7-Based Target Detection Network for Airborne Edge Computing

YOLOv7 marks a significant advancement in the YOLO series for object detection, offering substantial improvements in inference time and accuracy over other deep learning-based methods. This enhancement is particularly vital for high real-time airborne infrared detection systems. Despite the newer YOLOv9 [33], YOLOv7 provides faster processing speeds and adequate accuracy with fewer computational demands. It also introduces a more adaptable network structure and enhanced scalability, facilitating easier customization for specific application requirements by researchers.

YOLOv7 represents one of the more recent iterations within the YOLO series, renowned for its efficacy in object detection. The architecture of YOLOv7 is designed to process images in a single pass, thereby significantly reducing inference time while maintaining a high level of accuracy. The backbone of YOLOv7, which is crucial for feature extraction, leverages a compound scaling method that optimally balances the network's depth, width, and input resolution, enhancing the model's ability to capture and process complex features from input images. The detection net backbone is followed by the neck, which aggregates features at different scales, and the head, responsible for predicting bounding boxes and class probabilities.

As discussed above, we have innovatively enhanced the YOLOv7 model by serially incorporating a shallow-feature enhancement network (SE) before the detection net backbone for preprocessing and enhancing initial features of input images, and in parallel, integrating a deep-feature enhancement network (DE) to bolster the model's deep-feature analysis capabilities. These enhancements significantly improve the model's precision in object detection, catering especially to complex scenarios with subtle image features.

Furthermore, we adapt the YOLOv7 architecture for object detection, tailoring it to the constraints of computational and power resources characteristic of edge devices. As illustrated in Figure 4, our approach involves streamlining the network's backbone by omitting two convolutional layers from the original configuration. This refinement is aimed at reducing the computational burden, rendering the model more compatible with the limited processing capabilities of edge devices. Furthermore, we have simplified the original ELAN structure to a more compact version, termed L-ELAN, by reducing the number of ELANs from four to three. This modification efficiently trims the model's parameters, ensuring a lighter architecture while maintaining its detection performance.

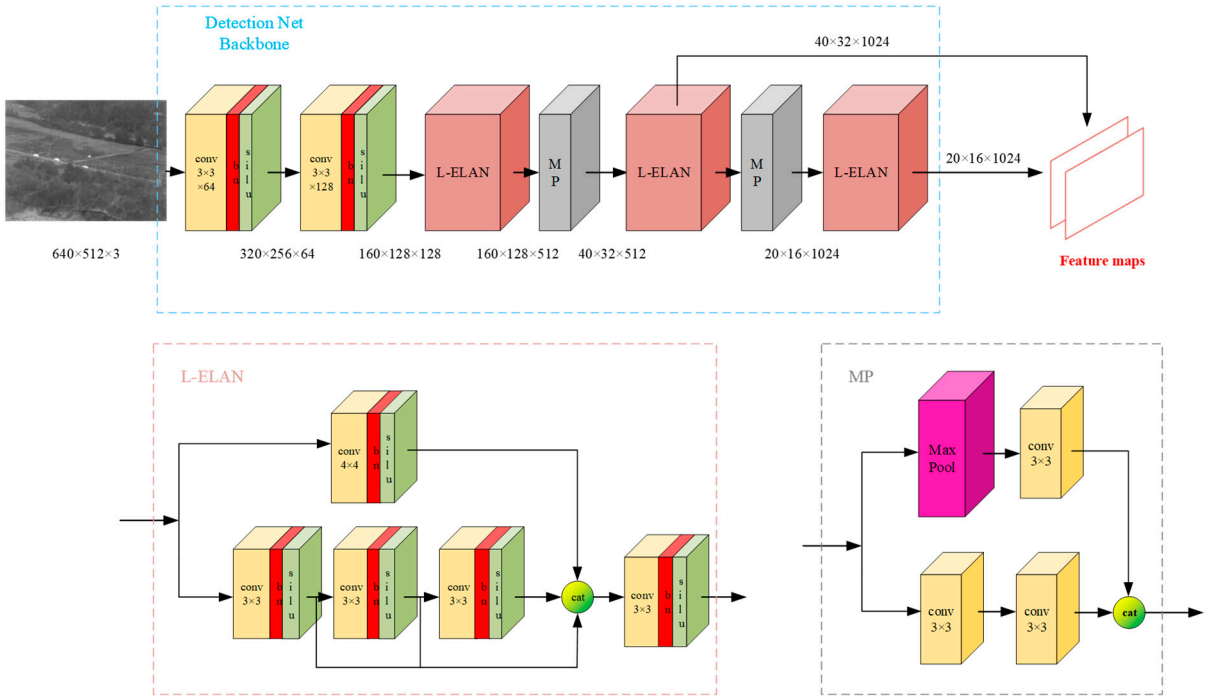


Figure 4. The backbone structure of YOLOv7 optimized for edge devices is illustrated, where the SiLU activation functions are denoted by green segments. Further details regarding the L-ELAN and MP (Max Pooling) structures are provided below.

2.2. Loss Function and Training Strategy

This section is divided into two parts: the first outlines the development of the loss function tailored for AIMED-Net, highlighting its significance in optimizing network performance; the second delves into the network's training strategy, detailing the methods employed to ensure efficient learning and accuracy improvement.

2.2.1. Feature Enhancement Loss and Object Detection Loss

In the shallow-feature enhancement network, a multi-scale loss function has been devised to thoroughly evaluate the discrepancies between low-quality images and their high-quality counterparts, focusing on both local and global features. This function integrates pixel difference loss, adversarial loss, similarity loss, and Total Variation Loss, each selected to comprehensively enhance the network's image enhancement capabilities.

1. Pixel Difference Loss

In the domain of infrared image enhancement networks, the reliance on the L2 loss for computing image differences predominantly focuses on the discrepancies between individual pixel values, thereby overlooking the relationships among pixels. This limitation results in a constrained perception of local features. To address this challenge, as delineated in Equation (1), we introduce a novel loss function that incorporates pixel relationships into

the computation of image discrepancies, enabling a more accurate assessment of image differences. This approach, by adjusting the weights $\omega_{(i,j)}$ based on the local similarity among pixels, effectively integrates the connectivity of regional pixels into the loss computation. High similarity between adjacent pixels results in increased weights, thereby emphasizing visually continuous or consistent areas within the loss function. This weighted mechanism not only reflects the differences between individual pixels but also facilitates a more comprehensive evaluation of the overall image discrepancies, especially highlighting areas that are visually continuous or exhibit high similarity.

$$L_{\text{pd}} = \frac{1}{CHW}\sum_{i,j}\omega_{(i,j)}|G(I)(i,j) - T(i,j)|^2 \tag{1}$$

where *C*, *H*, and *W* represent the count, height, and width of the images, respectively; *G(I)(i, j)* and *T(i, j)* respectively denote the pixel values of the enhanced image and the target high-quality image.

The calculation of $\omega_{(i,j)}$ is presented in Equation (2), where α is a tuning parameter used to control the extent to which similarity influences the weights. *D(i, j)* denotes the average difference of the surrounding 8 pixels at the corresponding positions of two images, as detailed in Equation (3).

$$\omega_{(i,j)} = e^{-\alpha \cdot D(i,j)} \tag{2}$$

$$D(i,j) = \frac{1}{8}\sum_{u=i-1}^{i+1}\sum_{v=j-1}^{j+1}|I_1(u,v) - I_2(u,v)|,\ (u \neq i,\ v \neq j) \tag{3}$$

2. Adversarial Loss

The discriminator extracts and recognizes texture information, performing binary classification on randomly input enhanced images and target high-quality images to determine the authenticity of the images. It outputs the estimation results. At the initial stage of training, considering the inferior performance of the generator, we opt to maximize $\log D(G(I))$ instead of minimizing $\log(1 - D(G(I)))$ in order to increase the gradient update for the generator's parameters, aiming to accelerate network training. The final adversarial loss function is defined as presented in Equation (4).

$$L_{\text{ad}} = -\sum_{i}\log D(G(I), T) \tag{4}$$

where *G* and *D* respectively denote the generator and discriminator networks, *I* represents the low-quality infrared image, $G(I)$ denotes the enhanced image, and *T* represents the target high-quality image, with *i* indicating the total number of training images. The discriminator networks and generator networks are jointly trained to minimize the adversarial loss.

3. Total Variation Loss

In alignment with the requirements for displaying infrared images, we incorporate Total Variation Loss into our loss function. The purpose is to eliminate noise introduced during the image enhancement process and to smooth the image to meet the visual perception needs of the human eye. The Total Variation Loss is presented as shown in Equation (5).

$$L_{\text{tv}} = \frac{1}{CHW}\left\|\nabla_x G(I) + \nabla_y G(I)\right\| \tag{5}$$

where *C*, *H*, and *W* represent the dimensions of the enhanced image $G(I)$.

4. Total Feature Enhancement Loss

The total infrared feature enhancement loss function designed by us is the weighted sum of the aforementioned three loss functions, as shown in Equation (6).

$$L_{eh} = \lambda_1 L_{pd} + \lambda_2 L_{\text{ad}} + \lambda_3 L_{tv} \tag{6}$$

where λ_1, λ_2, and λ_3 are weight parameters.

The loss function employed by the YOLO algorithm exhibits superior performance in global feature extraction and achieves real-time processing capabilities. However, given the unique attributes of airborne infrared imagery, including the presence of small-sized targets, low contrast, and environmental noise interference, it becomes apparent that the standard YOLO loss function does not adequately fulfill the requirements for applications within this specific context. Therefore, our research aims to refine the loss function of the YOLO algorithm for enhanced detection in airborne infrared scenarios, with a focus on significantly improving the detection capabilities for small and weakly visible targets.

1. CIoU Loss

In YOLO, the bounding box regression loss is defined as shown in Equation (7).

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \tag{7}$$

where $\rho^2(b, b^{gt})$ represents the Euclidean distance between the centers of the predicted and the actual bounding boxes, and c denotes the diagonal distance of the smallest rectangular area that can encompass both the predicted and the actual bounding boxes. α and v are the aspect ratios, with their formulas provided respectively in Equations (8) and (9).

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{8}$$

$$v = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2 \tag{9}$$

where w and h represent the height and width of the predicted bounding box, respectively, while w^{gt} and h^{gt} denote the height and width of the true bounding box, respectively.

To tackle the difficulty of detecting small targets in infrared imagery, we propose a scale-aware loss function designed to boost the network's capability in identifying faint targets. Traditional bounding box regression losses show limited sensitivity to small targets due to scale variations, leading to poorer detection performance. To address this, we incorporate a scale-aware factor into the conventional CIoU loss, inversely related to target size, thereby emphasizing the importance of small targets by increasing their weight in the total loss. This encourages the network to focus more on these harder-to-detect, smaller targets. Such an adjustment in weights enhances the detection precision of small targets while ensuring the detection of larger targets remains stable. This improvement enables our network to more fairly manage targets across scales, especially beneficial in infrared imaging scenarios characterized by low contrast and small targets, significantly boosting the detection's accuracy and robustness. The modified bounding box regression loss is detailed in Equation (10).

$$L_{sa-CIoU} = \sum_{i=1}^{N} \alpha_i \cdot L_{CIoU} \tag{10}$$

where α_i represents the scale-aware factor, as shown in Equation (11). w_i and h_i denote the length and width of the target bounding box, respectively. λ_s is a weight factor, and γ is a hyperparameter that controls the proportion of small targets.

$$\alpha_i = \lambda_s \left(\frac{1}{w_i \cdot h_i} \right)^{\gamma} \tag{11}$$

2. Objectness Loss

The objectness loss is defined as the degree of overlap between the ground truth boxes of positive samples and the predicted boxes, utilizing Binary Cross-Entropy (BCE) loss to quantify the discrepancy between the predicted values and actual values, as illustrated in Equation (12).

$$L_{obj} = -\frac{1}{N}\sum_{i=1}^{N}[t_i \log(c_i) + (1 - t_i)\log(1 - c_i)] \quad (12)$$

where t represents the Intersection over Union IoU value between the ground truth and predicted boxes, c denotes the predicted confidence, and N is the total number of images in a batch.

3. Classification Loss

The classification loss is defined as the similarity between the actual categories of positive samples and the predicted categories, also calculated using BCE loss, as demonstrated in Equation (13).

$$L_{cls} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \quad (13)$$

where y represents the true category labels, and p denotes the predicted category values.

4. Detection Loss

The detection loss is defined as the weighted sum of the three aforementioned losses, as shown in Equation (14), where λ_5, λ_6, and λ_7 represent the weight parameters. To ensure training stability, the initial values of these weights are set according to the default values of the YOLO and are subsequently adjusted through experimentation to determine the optimal values for specific scenarios.

$$L_{dt} = \lambda_5 L_{obj} + \lambda_6 L_{cls} + \lambda_7 L_{CIoU} \quad (14)$$

2.2.2. Model Training Strategy

To ensure optimal performance of the network in target detection, we commence with pre-training specific components of the network prior to joint training. The components that necessitate pre-training encompass the shallow-feature enhancement network and the deep-feature enhancement network.

1. Pre-training of the Shallow-Feature Enhancement Network

The shallow-feature enhancement network is pre-trained by optimizing the network weights through backpropagation using the total feature enhancement loss function (Equation (6)). Subsequently, the resulting pre-trained shallow-feature enhancement network model is saved.

2. Pre-training of the Deep-Feature Enhancement Network

To address the issue of task separation between the image enhancement network and the target detection network, which hinders effective parameter optimization of the image enhancement network toward improving detection outcomes, we employ the total detection loss function (Equation (14)) for backpropagation. This optimization is facilitated by the backbone and head sections of the Detection Net (Figure 3). Ultimately, we save the pre-trained weights of the deep-feature enhancement network, resulting in a pre-trained model.

3. Joint Training and Optimization

Finally, joint training of the AIMED-Net (Figure 1) is performed, where both the shallow- and deep-feature enhancement networks leverage pre-trained weights. The joint training employs the total detection loss function (Equation (14)) for backpropagation to optimize the overall network's weight parameters. Ultimately, the trained weights are saved to generate the model.

2.3. Dataset

The HIT-UAV dataset (Figure 5) is a comprehensive high-altitude infrared thermal imaging dataset specifically designed for drone target detection applications. It consists

of 2898 infrared thermal images with a resolution of 640 × 512, extracted from a vast collection of 43,470 frames captured by drones in various scenarios including schools, parking lots, roads, and playgrounds. Due to its extensive coverage and public availability, we have chosen the HIT-UAV dataset as our base dataset. We then constructed an enhanced dataset based on this foundation, which was subsequently used for ablation studies and comparative experiments.



Figure 5. The HIT-UAV dataset.

Infrared imaging devices, due to inherent defects, generate both fixed and random noise during the process of image acquisition [34], as depicted in Figure 6. The fixed noise arises from the non-uniform response rate of the infrared detector itself, imaging defects, and clutter interference. It is mathematically represented as multiplicative noise according to Equation (15).

$$x_i = a_i I + b_i \qquad (i = 1, 2, \ldots, M) \tag{15}$$

where, I represents the incident infrared radiation, x_i denotes the response of the detector element, and M is the number of detector elements in the array. The gain (a_i) and offset (b_i) of each element reflect the non-uniformity of the individual detectors.



Figure 6. Infrared images typically exhibit both fixed and random noise.

The random noise observed in infrared systems primarily arises from fluctuations in background radiation, photoelectric conversion within detectors, and additional noise introduced by signal readout and processing circuits. This noise can be accurately characterized by Gaussian and Poisson distributions, with their respective probability density functions (PDFs) presented in Equations (16) and (17).

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \tag{16}$$

where x denotes the grayscale value of the imaging noise; μ represents the mean or expected value of x; and σ^2 denotes the variance of x.

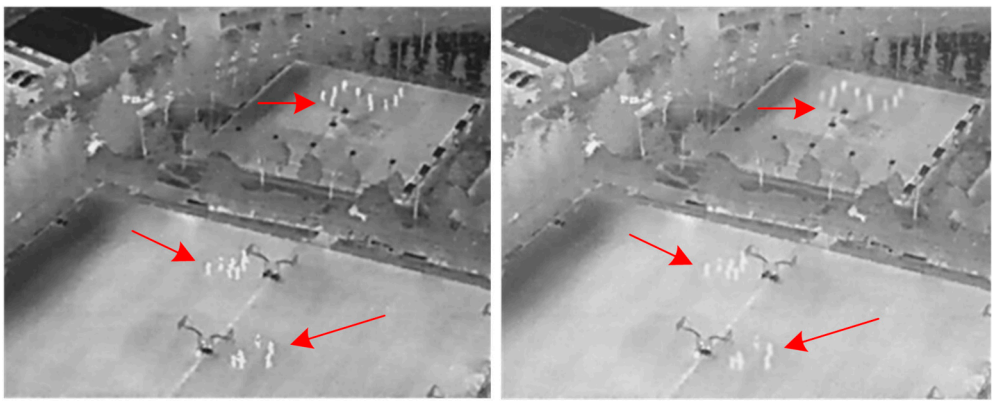
$$P(x = m) = \frac{\lambda^m e^{-\lambda}}{m!} \qquad (m = 0, 1, 2, \ldots; \lambda > 0) \tag{17}$$

Noise interference degrades image details, lowers contrast, and diminishes visual effects in infrared imaging, challenging target detection. The detection of small targets is particularly challenging; their limited size and weak signal, compounded by high noise levels, increase susceptibility to obscuration. Consequently, this complicates the accurate identification and tracking of small targets for detection algorithms.

In image enhancement tasks, a specific ground truth (GT) is often unavailable. To tackle this challenge, we integrate specific infrared noise into designated areas of the original images. These modified images are utilized as inputs, with the unaltered originals acting as the GT for network training. Through comparing the modified inputs with the original images, the network acquires the ability to restore images and enhance target features. Following this, the original images undergo iterative processing through the network to produce images of higher quality with enhanced features, thus enhancing the performance of the target detection network.

To construct high- and low-quality image pairs for training, specific small target regions were manually annotated on the original images. Subsequently, these regions were subjected to the addition of Gaussian noise, Poisson noise, and stripe noise at various random intensities to generate degraded images.

The dataset utilized for the pre-training of the shallow-feature enhancement network targeting infrared images was independently developed, drawing upon the HIT-UAV dataset. It consists of 2898 pairs of original infrared images alongside their corresponding images, which have been artificially degraded by introducing simulated infrared noise in designated areas, as detailed in Figure 7. Out of these, 2318 pairs were designated as the training set, with the remaining 580 images allocated as the validation set.



(**a**) (**b**)

Figure 7. Comparison between the original images (**a**) and the degraded infrared images (**b**) generated by simulating infrared noise. The red arrows indicate the specific areas where the aforementioned noise is increased.

We employed the original HIT-UAV dataset as the dataset for our target detection network. The HIT-UAV dataset encompasses 24,899 labels across four categories: Person, Car, Bicycle, and Other Vehicle. For our dataset configuration, we allocated 1622 images to the training set, 580 images to the test set, and 696 images to the validation set. As indicated in Table 1, the final dataset comprises 17,118 labels for small targets of less than 32×32 pixels, 7249 labels for medium targets of less than 96×96 pixels, and 384 labels for large targets. Notably, the smallest targets in the HIT-UAV dataset occupy only 0.01% of the image pixels, meeting the criteria for detecting small aerial infrared targets using drones.

Table 1. Annotation counts for small, medium, and large targets in the HIT-UAV dataset.

	Small	**Medium**	**Large**
	(0, 32 × 32)	**(32 × 32, 96 × 96)**	**(96 × 96, 640 × 512)**
HIT-UAV	17,118	7249	384
Train set	12,045	5205	268
Test set	3331	1379	70
Validation set	1742	665	46

2.4. Edge Deployment Solution Based on Vitis-AI

In the full-stack AI deployment framework for edge computing devices, Vitis-AI serves as both the compiler and backend, receiving network parameters trained by the front-end DNN (Deep Neural Network) framework. It then optimizes and compiles these parameters, and passes them to the backend for invocation by edge devices. The algorithm deployment process is illustrated in Figure 8. On the host side, the PyTorch deep learning framework is initially used to construct and train the base model. Subsequently, the Vitis-AI tool is employed for model pruning to obtain an optimized model, which is further quantized—converting floating-point weights to fixed-point representation to accommodate the computational characteristics of FPGAs. Finally, the Vitis-AI compiler compiles the model to generate compiled files. Moreover, the compiler optimizes the network model for specific Xilinx hardware architectures, such as the ZU5EV chip, to achieve optimal performance and resource utilization. On the edge side, the optimized model can be loaded and run on the Zynq UltraScale+ MPSoC ZU5EV chip.

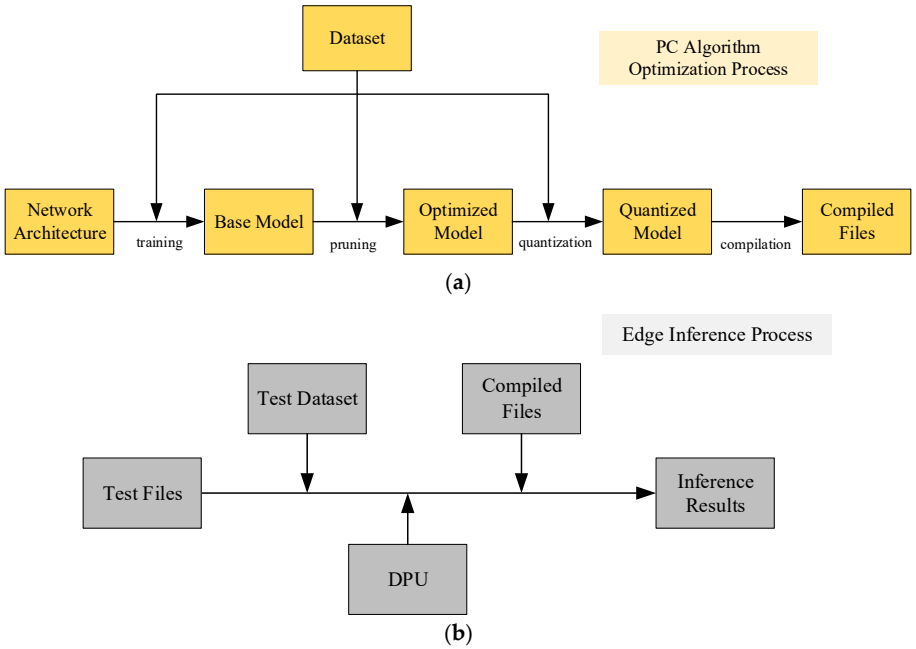


Figure 8. Optimization and deployment process for deep neural network models using the Vitis-AI, with (**a**) representing the host-side optimization process and (**b**) denoting the edge-side inference process.

2.5. Assessment Indicators

In this study, we initially assess the performance of the shallow-feature enhancement network in isolation. The quality of the enhanced images is evaluated utilizing Image Entropy (IE) and Mean Gradient (MG), as delineated in Equations (18) and (19), respectively. IE quantitatively reflects the amount of information contained within an image. Notably, in scenarios involving low-quality infrared small targets, lower entropy values suggest a reduction in background noise, which predominantly constitutes the image, thereby accentuating the information pertaining to small targets. Conversely, MG is employed as a metric for assessing the sharpness and texture variation within an image, where a higher MG signifies improved image detail and clarity.

$$IE = -\sum_{i=0}^{L-1} p(i) \log_2 p(i) \tag{18}$$

where L represents the grayscale levels of the image, while $p(i)$ denotes the probability of occurrence of the grayscale level ii within the image. This probability can be calculated by dividing the frequency of each grayscale level by the total number of pixels in the image.

$$MG = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} \sqrt{(\Delta x)^2 + (\Delta y)^2} \tag{19}$$

where M and N denote the size of the image, while Δx and Δy represent the pixel difference in the horizontal and vertical directions, respectively.

In terms of overall detection performance, we adopt a comprehensive suite of evaluation metrics. These include the F1 score, Average Precision (AP), Mean Average Precision (mAP), the number of parameters, GFLOPs, and the inference time per frame on edge devices for convolutional computations. The F1 score is computed as the weighted average of precision and recall, as specified in Equations (20) and (21). Both AP and mAP are utilized as definitive indicators of model detection accuracy. Additionally, the model's size and algorithmic complexity are quantified through the number of parameters and GFLOPs, respectively. The inference time on edge devices reflects the operational speed of the model in edge computing environments. Detailed formulations for these evaluation parameters are provided in Equations (22)–(24).

$$Precision = \frac{TP}{TP + FP} \tag{20}$$

$$Recall = \frac{TP}{TP + FN} \tag{21}$$

$$F1 = 2 * \frac{P * R}{P + R} \tag{22}$$

$$AP = \int_0^1 P(r) dr \tag{23}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{24}$$

The terms True Positives (TP), False Positives (FP), and False Negatives (FN) are commonly used to denote samples that are correctly identified, incorrectly identified, and missed, respectively. N is a variable representing the number of classes being categorized.

3. Results

To evaluate the effectiveness of the approach we propose, ablation experiments were conducted. This section is divided into two parts: initially, a comprehensive description of

our experimental platform is provided; subsequently, ablation analyses are carried out to assess the efficacy of each component within the proposed method.

3.1. Experimental Platform and Parameter Settings

The hardware platform settings used in the experiments of this article are shown in Table 2. Our experiments were conducted under the Python 3.10.12 torch-2.1.1+cu121 framework, utilizing a computing platform equipped with an NVIDIA GeForce RTX 3090 GPU for both training and testing phases. The model was quantized and compiled through the Vitis-AI tool and edge deployment and inference were completed on the Xilinx Zynq UltraScale+ MPSOC FPGA.

Table 2. Experimental platform and details of this article.

Names	Related Configurations
Graphics processing unit	NVIDIA GeForce RTX 3090
Central processing unit	Inter(R) Core™ i9-13900K
GPU memory size	24 G
Operating system	Ubuntu 22.04
Computing platform	CUDA 12.1
Deep learning framework	Pytorch 2.1.1
Quantization compiler tool	Vitis-AI 3.0
Edge computing inference platform	Zynq UltraScale + MPSoC ZU5EV

3.2. Ablation Studies and Analysis

To assess the effectiveness of the proposed model, a series of ablation studies were carried out on the HIT-UAV dataset. These studies aimed to determine the influence of individual components on the overall performance of the integrated network.

3.2.1. Analysis of the Shallow-Feature Enhancement Network

The visual effects of the shallow-feature enhancement network are validated by Figure 9. Panel (a) presents the original image, while panel (b) exhibits the feature-enhanced image. It is apparent that the suppression of infrared noise and the accentuation of detailed contours of small targets have been effectively achieved. To further quantitatively analyze the impact of the shallow-feature enhancement network on overall performance, 1000 images were randomly selected from the dataset to compare their average IE and MG before and after enhancement, as shown in Table 3. The decrease in IE and the increase in MG post-enhancement indicate that the shallow-feature enhancement network effectively accentuates the feature information of small targets.

Table 3. Comparison of Image Quality Metrics with and without Shallow-Feature Enhancement.

Shallow-Feature Enhancement	IE	MG
	6.90	2.98
√	6.58	3.51

3.2.2. Analysis of the Deep-Feature Enhancement Network

The network visualization analysis was conducted using Grad-CAM (Gradient-weighted Class Activation Mapping). As illustrated in Figure 10, panel (a) demonstrates the feature heatmap output obtained from the fourth-to-last layer of the Detection Net Head without incorporating the deep-feature enhancement network structure. Panel (b) exhibits the output at the same layer after integrating the deep-feature enhancement network. It is evident that incorporating the deep-feature enhancement network not only improves the detection network's focus on small targets but also reduces its attention toward background noise. A quantitative analysis of the enhancement effects of the deep-feature enhancement network will be provided in Section 3.2.3.

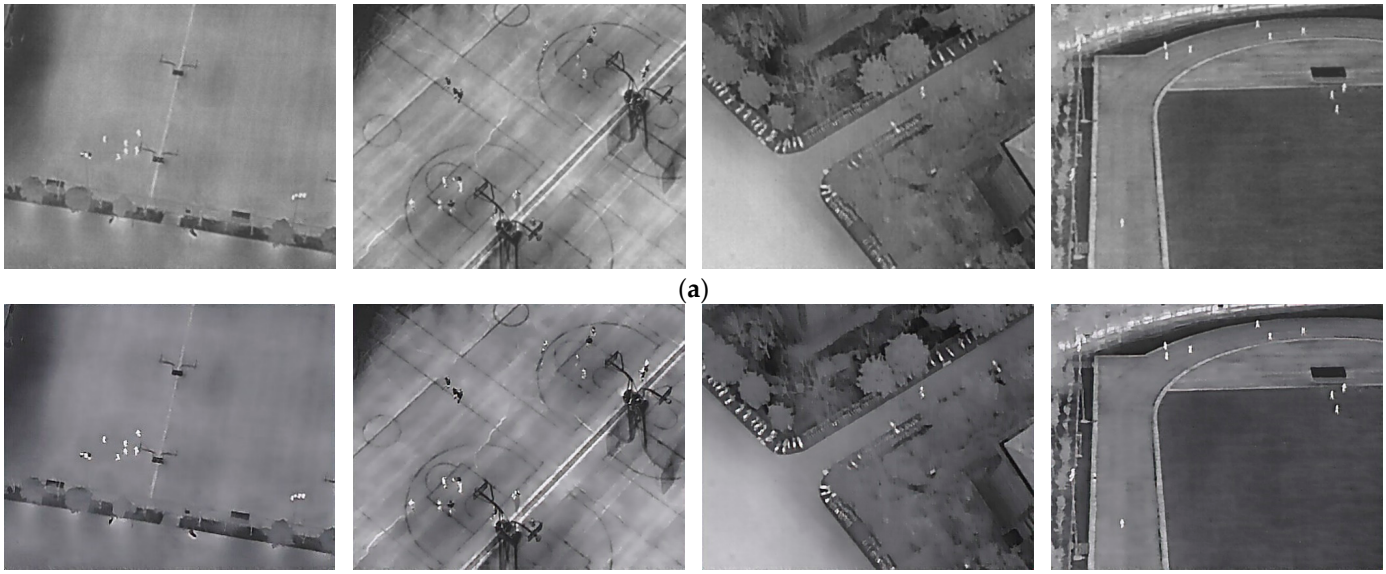


Figure 9. Comparison of the original image with the image enhanced by the shallow-feature enhancement network. Panel (**a**) displays the original image, while panel (**b**) shows the enhanced image.

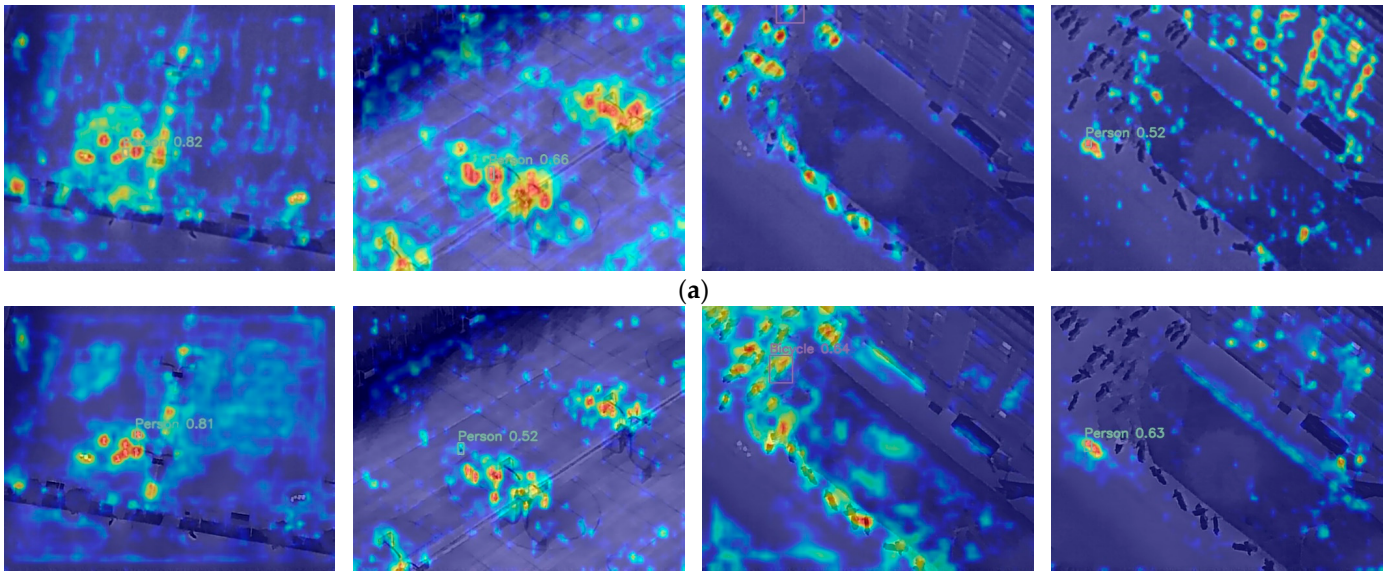


Figure 10. Comparative attention heatmap of the penultimate layer outputs from the Detection Net Head before and after enhancement by the deep-feature enhancement network is presented. (**a**) shows the attention heatmap for target detection without deep-feature enhancement, while (**b**) illustrates the attention heatmap for the same image after applying deep-feature enhancement. Regions highlighted in red indicate regions of interest to the target detection network, while areas shaded in blue represent regions that receive comparatively lower attention.

3.2.3. AIMED-Net Ablation Analysis

Ablation experiments were conducted using an input image size of 640 × 640 and a batch size of 16, and each network was trained for 300 epochs. Detailed results are presented in Table 4. When utilizing only the lightweight modified YOLOv7s backbone structure, despite a reduction of 12.4% in parameter count, there was a decrease of 3.9% in mAP performance. However, incorporating the shallow-feature enhancement network led to notable improvements in AP values for small targets such as Person (2.3% increase)

and overall mAP (4.7% improvement), without significantly increasing the number of parameters employed. Furthermore, by including the deep-feature enhancement network, additional enhancements were observed for difficult-to-distinguish categories like OtherVehicle through improved representation of the infrared target's deep features, resulting in an extra increase of 2.8%, building upon the benefits provided by the shallow-feature enhancement network alone. Moreover, replacing the traditional CIoU loss function with our proposed sa-CIoU loss function further enhanced AP values for Person and OtherVehicle categories by increments of 0.5% and 0.4%, respectively; this can be attributed to increased weightage given to small targets within the total loss while maintaining stability in detecting large-sized targets.

Table 4. Ablation Experiments Based on the HIT-UAV Dataset.

YOLOv7s	L-Backbone	SE	DE	sa-CIoU	Parameters	F1 (%)	AP_{50}				mAP (%)	Time (ms)
							AP_{Person} (%)	AP_{Car} (%)	$AP_{Bicycle}$ (%)	AP_{Other} (%)		
√					37.2 M	84.5	87.5	95.1	85.6	71.5	84.9	7.9
√	√				32.6 M	83.6	85.5	93.2	82.1	63.2	81.0	6.3
√	√	√			32.7 M	85.9	87.7	94.5	86.4	74.4	85.7	6.5
√	√		√		34.1 M	85.7	86.8	93.4	85.5	76.8	85.6	6.6
√	√			√	32.6 M	83.9	86.6	93.8	84.3	64.5	82.3	6.3
√	√	√	√		34.2 M	87.1	88.2	96.1	87.5	77.2	87.2	6.7
√	√	√	√	√	34.2 M	87.4	88.7	96.2	87.6	77.6	87.5	6.7

The experimental results are illustrated in Figure 11. It is observed that the detection performance of YOLOv7s diminishes after lightweight optimization. Following the sole addition of the shallow-feature enhancement network (SE), there is a notable improvement in both the visual quality of images and the detection performance of small targets, although false positives persist. The incorporation of the deep-feature enhancement network (DE) alone enhances the deep features of targets, leading to a reduction in false positives. The integration of both SE and DE mitigates issues of missed and false detections. Further enhancement of the network's ability to perceive small targets is achieved by adding an optimized sa-CIoU loss function.

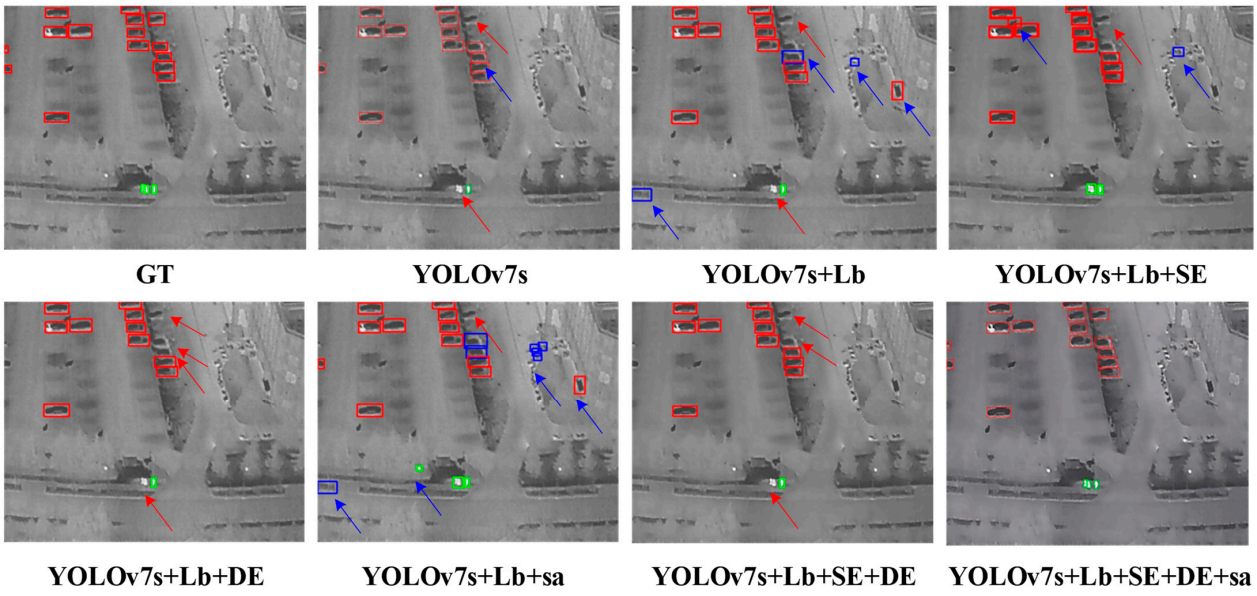


Figure 11. The ablation analysis of different modules is presented, with instances of missed detections highlighted by red arrows and instances of false detections denoted by blue arrows. Different colors of detection boxes indicate different classes.

The performance comparison chart in Figure 12 demonstrates the superiority of the integrated network for multi-layer feature enhancement and detection of infrared small targets in the ablation study. Therefore, our proposed method provides a more suitable solution for airborne edge computing scenarios involving infrared target detection tasks.

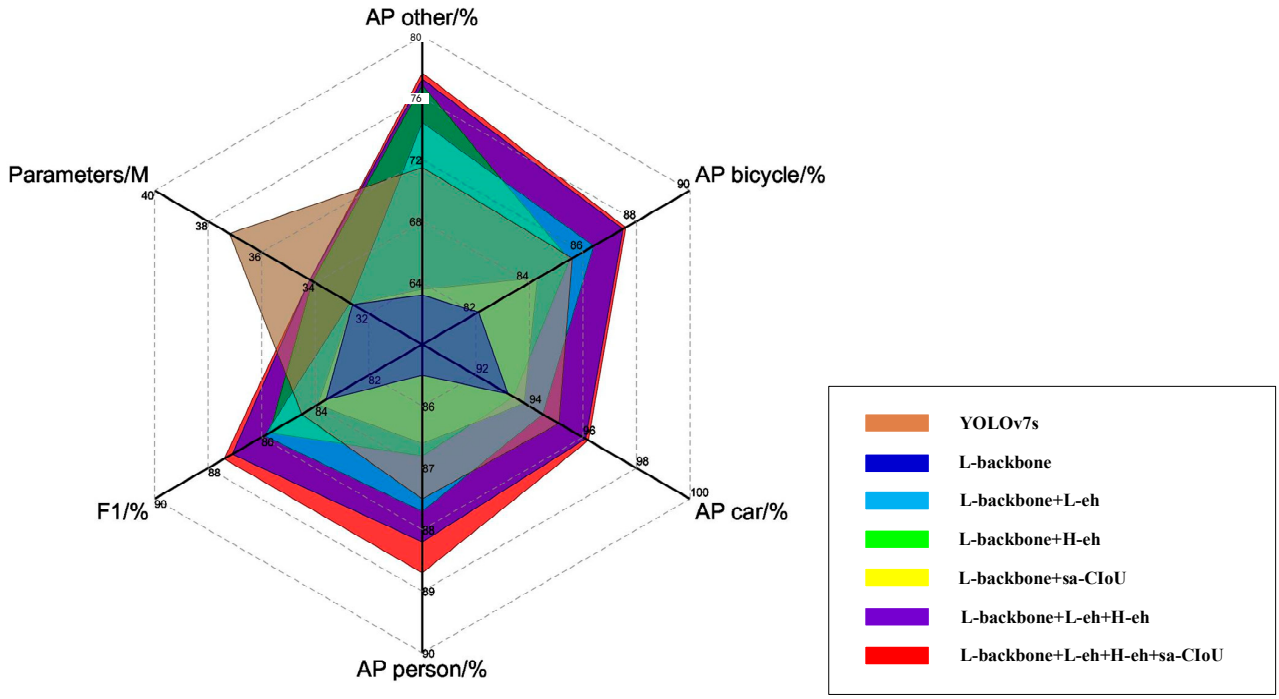


Figure 12. Comparative evaluation of model performance in ablation experiments.

4. Discussion

To further validate the reliability of our method, we selected other models from the YOLO series, including the latest YOLOv9 model, as baselines for experiments conducted on the HIT-UAV Unmanned Aerial Vehicle Infrared Target Dataset. This dataset captures targets at altitudes ranging from 60 to 130 m, resulting in complex and variable image backgrounds and target scales. Moreover, the majority of the targets are small, presenting substantial challenges for detection. The results in Table 5 demonstrate our method's significant advantages in detecting infrared targets in complex airborne edge environments.

Table 5. Performance comparison with advanced algorithms.

Model	Parameters	F1 (%)	AP_{50}				mAP (%)	Time (ms)
			AP_{Person} (%)	AP_{Car} (%)	$AP_{Bicycle}$ (%)	AP_{Other} (%)		
YOLOv4 [35]	52.5 M	78.9	79.1	91.8	70.8	68.1	77.5	14.4
YOLOv5s	9.1 M	84.8	92.6	97.8	90.3	69.4	87.5	4.4
YOLOv7s	37.2 M	84.5	87.5	95.1	85.6	71.5	84.9	7.9
YOLOv8s [36]	11.2 M	85.3	92.3	97.1	89.9	71.9	87.8	5.3
YOLOv8l	43.7 M	86.1	91.6	97.6	90.3	68.1	86.9	8.4
YOLOv9	60.8 M	89.3	92.2	97.3	92.7	88.0	92.6	16.8
AIMED-Net	34.2 M	87.4	88.7	96.2	87.6	77.6	87.5	6.7

The edge inference time of our enhanced detection integrated network model increased by 52.3% and 26.4% compared to YOLOv5s and YOLO8s, respectively, at an input resolution of 640 × 640. Despite this increase, real-time processing was still achieved, meeting the application requirements of airborne platforms. Furthermore, the F1 score improved by 2.6% and 2.1% compared to YOLOv5s and YOLO8s, respectively. In scenarios where the AP levels for other categories were similar, there was an improvement of 8.2% and 5.7%, respectively, in the AP for the challenging-to-detect class OtherVehicle category due to our model's shallow- and deep-feature enhancement network structures that enhance target features in complex environments, thereby improving robustness and detection accuracy significantly. Notably, our enhanced detection integrated network reduced model complexity by 46.2% compared to the current state-of-the-art algorithm YOLOv9 while also shortening inference time by 60%. This reduction ensures low power consumption along with high real-time processing capabilities on edge devices, which is advantageous for drones' quick response times as well as prolonged endurance.

The outcomes of various algorithms are depicted in Figure 13, showcasing the significant progress our method has achieved in the field of drone-based infrared target detection. Confronted with the task of detecting diminutive targets, such as individuals within complex environments, simpler network architectures like YOLOv5s and YOLOv8s struggle to accurately identify these targets. This inefficacy results in a combination of missed detections and false positives. Conversely, more sophisticated networks such as YOLOv8l and YOLOv9 demonstrate enhanced proficiency in identifying small targets within intricate settings. Consequently, they exhibit a reduction in both missed detections and false positives, alongside achieving a superior mAP. It is noteworthy that all models, with the exception of AIMED-Net and YOLOv8l, faced challenges in the accurate classification of objects within the OtherVehicle category. This issue stems from the subtle differences in features between the OtherVehicle and Car categories, which are challenging to discern in low-resolution infrared imagery, leading to incorrect identifications as Cars by the networks.

These experimental findings demonstrate that our integrated network for multi-layer enhancement of infrared features alongside small target detection enhances feature prominence for targets within complex scenes. Thus, it improves overall performance without introducing excessive computational overhead. Such advantages make it suitable for practical applications within airborne edge computing environments.

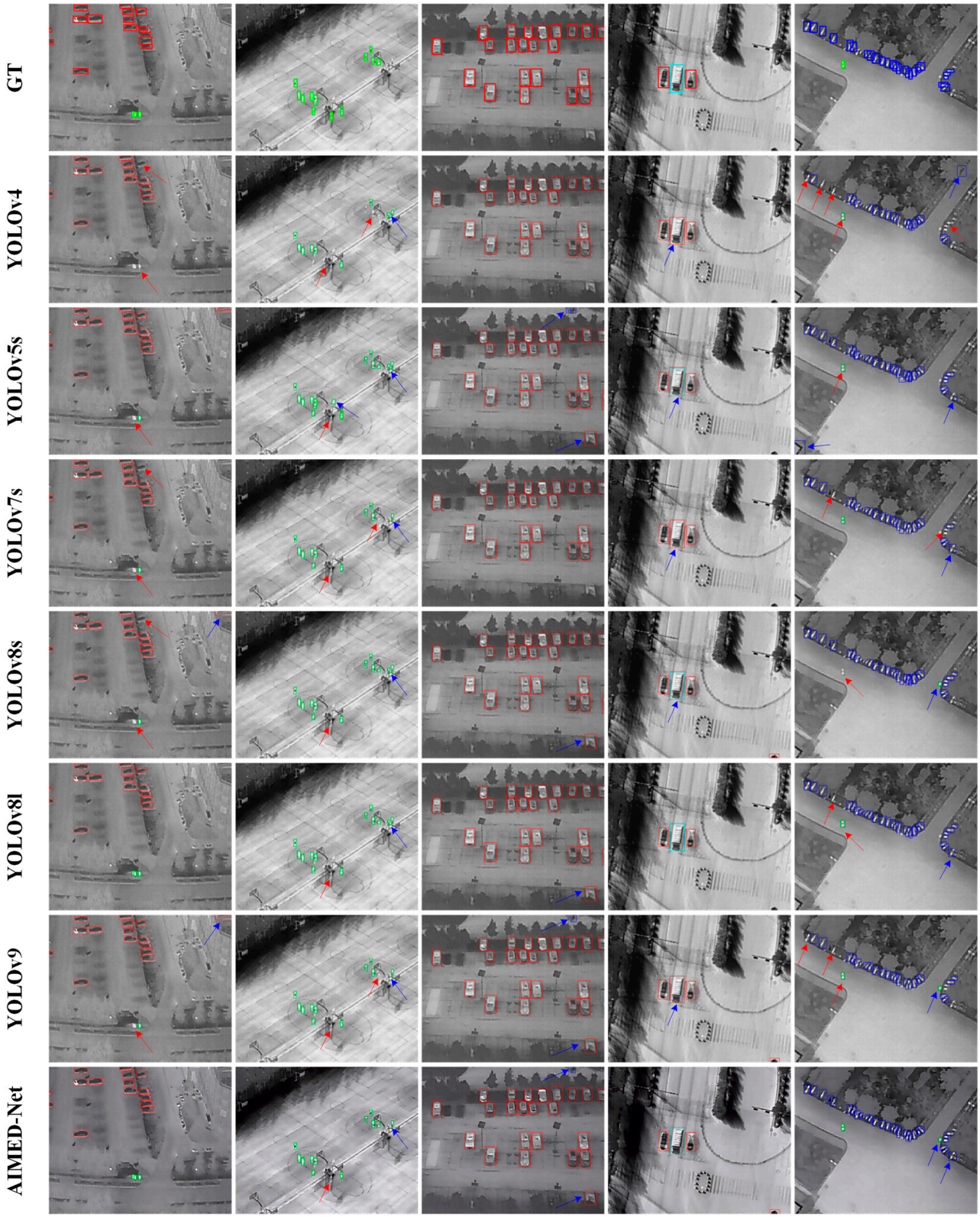


Figure 13. Comparative outcomes obtained using advanced algorithms are presented, with instances of missed detections highlighted by red arrows and instances of false detections denoted by blue arrows. Different colors of detection boxes indicate different classes.

5. Conclusions

This article introduces AIMED-Net, an enhancing infrared small target detection net in UAVs with multi-layer feature enhancement for edge computing. The proposed approach consists of a shallow-feature enhancement network based on generative adversarial principles and a deep-feature enhancement network focused on detection. To improve target detection performance through multi-scale infrared feature enhancement, we propose an infrared image-feature enhancement method that incorporates modeled infrared image noise into original images to create high- and low-quality image pairs. By leveraging the YOLOv7 framework as a foundation, we have developed an enhanced target detection network that integrates multiple feature enhancement techniques. This network is specifically optimized for detecting infrared targets in edge computing environments, reducing model complexity while enhancing robustness and accuracy in recognizing small targets.

Experimental results using the HIT-UAV public dataset demonstrate that our improved approach achieves significant improvements compared to the original YOLOv7s model. Specifically, it shows a 2.5% increase in F1 score, a 2.6% increase in mAP (mean average precision), and a 6.1% increase in AP (average precision) for detecting OtherVehicle targets. Additionally, our approach reduces model parameters by 5.4%, computational workload by 29.9%, and inference time at the edge by 15.2%. Compared with existing state-of-the-art methods, our approach strikes a balance between detection efficiency and accuracy, making it suitable for practical applications in aerial edge environments.

Our approach constitutes a significant advancement in the realm of infrared small target detection for unmanned aerial vehicles (UAVs), acknowledging, however, that it is not without its limitations and areas ripe for further exploration. The current iteration of AIMED-Net, despite its innovative integration of deep learning methodologies and FPGA edge devices to augment detection capabilities, highlights the potential for further enhancements, particularly in terms of model lightweighting and the acceleration of hardware performance. Future versions should aim to incorporate more nuanced optimizations designed to boost hardware efficiency and detection speeds. Such improvements could significantly extend the practical applicability and effectiveness of AIMED-Net in the demanding contexts of real-time edge computing environments.

Author Contributions: Conceptualization, T.L. and L.P.; methodology, L.P.; software, L.P.; validation, L.P., B.C. and Y.C.; formal analysis, J.C.; investigation, B.C.; resources, J.C.; data curation, B.C.; writing—original draft preparation, L.P.; writing—review and editing, T.L.; visualization, L.P.; supervision, J.C.; project administration, J.C.; funding acquisition, J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Hunan Province, China, grant number 2020JJ4670.

Data Availability Statement: Data related to this study are available from the corresponding author upon reasonable request. The codes used during this study are available from the corresponding author upon request. The data are not publicly available due to the specific policies of the institution.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang, S.; Liu, F. Infrared and visible image fusion based on non-subsampled shearlet transform, regional energy, and co-occurrence filtering. *Electron. Lett.* **2020**, *56*, 761–764. [CrossRef]
2. Lin, T.; Maire, M.; Belongie, S. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
3. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 23–28 June 2014.
4. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
5. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]

6. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot Multibox Detector. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016; Springer: Cham, Switzerland, 2016; pp. 21–37.
7. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
8. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
9. Liu, S.; Ma, Z.; Chen, B. Remote Sensing Image Detection Based on FasterRCNN. In *Artificial Intelligence in China*; Springer: Berlin/Heidelberg, Germany, 2021.
10. Wu, C.; Zhao, B.; Tang, L.; Zhao, B. Small Vehicles Detection Based on UAV. *J. Eng.* **2019**, *21*, 7894–7897.
11. Liu, M.; Du, H.Y.; Zhao, Y.J.; Dong, L.Q.; Hui, M.; Wang, S.X. Image Small Target Detection Based on Deep Learning with SNR Controlled Sample Generation. *Curr. Trends Comput. Sci. Mech. Autom.* **2017**, *1*, 211–220.
12. McIntosh, B.; Venkataramanan, S.; Mahalanobis, A. Infrared Target Detection in Cluttered Environments by Maximization of a Target to Clutter Ratio (TCR) Metric Using a Convolutional Neural Network. *IEEE Trans. Aerosp. Electron. Syst.* **2020**, *57*, 485–496. [CrossRef]
13. Park, J.; Chen, J.; Cho, Y.K.; Kang, D.Y.; Son, B.J. CNN-Based Person Detection Using Infrared Images for Night-Time Intrusion Warning Systems. *Sensors* **2019**, *20*, 34. [CrossRef]
14. Hou, Q.; Wang, Z.; Tan, F.; Zhao, Y.; Zheng, H.; Zhang, W. RISTDnet: Robust Infrared Small Target Detection Network. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
15. Zhao, B.; Wang, C.; Fu, Q.; Han, Z. A Novel Pattern for Infrared Small Target Detection with Generative Adversarial Network. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4481–4492. [CrossRef]
16. Zhang, T.; Peng, Z.; Wu, H.; He, Y.; Li, C.; Yang, C. Infrared Small Target Detection via Self-Regularized Weighted Sparse Model. *Neurocomputing* **2021**, *420*, 124–148. [CrossRef]
17. Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Asymmetric Contextual Modulation for Infrared Small Target Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 950–959.
18. Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Attentional Local Contrast Networks for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 9813–9824. [CrossRef]
19. Li, R.; Shen, Y. YOLOSR-IST: A Deep Learning Method for Small Target Detection in Infrared Remote Sensing Images Based on Super-Resolution and YOLO. *Signal Process.* **2023**, *208*, 108962. [CrossRef]
20. Liu, F.; Gao, C.; Chen, F.; Meng, D.; Zuo, W.; Gao, X. Infrared Small and Dim Target Detection with Transformer under Complex Backgrounds. *IEEE Trans. Image Process.* **2023**, *32*, 5921–5932. [CrossRef]
21. Zhang, T.; Li, L.; Cao, S.; Pu, T.; Peng, Z. Attention-Guided Pyramid Context Networks for Detecting Infrared Small Target Under Complex Background. *IEEE Trans. Aerosp. Electron. Syst.* **2023**, *59*, 4250–4261. [CrossRef]
22. Ye, Z.; Cai, Z. Non-salient Object Detection Algorithm Based on Infrared Image. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1002005. [CrossRef]
23. Guo, H.; Lu, T.; Wu, Y. Dynamic Low-Light Image Enhancement for Object Detection via End-to-End Training. In Proceedings of the 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 5611–5618.
24. Liu, H.; Jin, F.; Zeng, H.; Pu, H.; Fan, B. Image Enhancement Guided Object Detection in Visually Degraded Scenes. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *1*, 1–14. [CrossRef]
25. Liu, S.; Chen, P.; Woźniak, M. Image Enhancement-Based Detection with Small Infrared Targets. *Remote Sens.* **2022**, *14*, 3232. [CrossRef]
26. Huang, S.; Le, T.; Jaw, D. DSNet: Joint Semantic Learning for Object Detection in Inclement Weather Conditions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 2623–2633. [CrossRef]
27. Chen, L.; Jiang, Z.; Tong, L.; Liu, Z.; Zhao, A.; Zhang, Q.; Dong, J.; Zhou, H. Perceptual Underwater Image Enhancement with Deep Learning and Physical Priors. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 3078–3092. [CrossRef]
28. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023.
29. Suo, J.; Wang, T.; Zhang, X.; Chen, H.; Zhou, W.; Shi, W. HIT-UAV: A High-Altitude Infrared Thermal Dataset for Unmanned Aerial Vehicle-Based Object Detection. *Sci. Data* **2023**, *10*, 227. [CrossRef]
30. Kuang, X.; Sui, X.; Liu, Y.; Chen, Q.; Gu, G. Single Infrared Image Enhancement Using a Deep Convolutional Neural Network. *Neurocomputing* **2019**, *332*, 119–128. [CrossRef]
31. Johnson, J.; Alahi, A.; Li, F. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In Proceedings of the European Conference on Computer Vision, Cham, Switzerland, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 694–711.
32. Aly, H.A.; Dubois, E. Image Up-Sampling Using Total-Variation Regularization with a New Observation Model. *IEEE Trans. Image Process.* **2005**, *14*, 1647–1659. [CrossRef]
33. Wang, C.Y.; Yeh, I.H.; Liao, H.Y.M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. *arXiv* **2024**, arXiv:2402.13616.
34. Li, M.; Nong, S.; Nie, T.; Han, C.; Huang, L.; Qu, L. A Novel Stripe Noise Removal Model for Infrared Images. *Sensors* **2022**, *22*, 2971. [CrossRef]

35. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
36. Reis, D.; Kupec, J.; Hong, J.; Daoudi, A. Real-Time Flying Object Detection with YOLOv8. *arXiv* **2023**, arXiv:2305.09972.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.