drones

MDPI

Article

YOLO-SMUG: An Efficient and Lightweight Infrared Object Detection Model for Unmanned Aerial Vehicles

Xinzhe Luo [1] and Xiaogang Zhu [2,*]

[1] Jiluan Academy, Nanchang University, Nanchang 330031, China; 9107122046@email.ncu.edu.cn
[2] Big Data and Cyber Security Research Institute, Nanchang University, Nanchang 330031, China
* Correspondence: ncuzxg@ncu.edu.cn

Abstract: To tackle the high computational demands and accuracy limitations in UAV-based infrared object detection, this study proposes YOLO-SMUG, a lightweight detection algorithm optimized for small object identification. The model incorporates an enhanced backbone architecture that integrates the lightweight Shuffle_Block algorithm and the Multi-Scale Dilated Attention (MSDA) mechanism, enabling effective small object feature extraction while significantly reducing parameter size and computational cost without compromising detection accuracy. Additionally, a lightweight inverted bottleneck structure, C2f_UIB, along with the GhostConv module, replaces the conventional C2f and standard convolutional layers. This modification decreases computational complexity while maintaining the model's ability to capture and integrate essential feature information across multiple scales. Furthermore, the standard CIoU loss is substituted with MPDIoU loss, improving object localization accuracy and enhancing overall positioning precision in infrared imagery. Experimental results on the HIT-UAV dataset, which consists of infrared imagery collected by UAV platforms, demonstrate that YOLO-SMUG outperforms the baseline YOLOv8s, achieving a 3.58% increase in accuracy, a 6.49% improvement in the F1-score, a 57.04% reduction in computational cost, and a 64.38% decrease in parameter count. These findings underscore the efficiency and effectiveness of YOLO-SMUG, making it a promising solution for UAV-based infrared small object detection in complex environments.

Keywords: object detection; YOLOv8s; lightweight networks; UAV; infrared image

Academic Editor: Pablo Rodríguez-Gonzálvez

Received: 19 February 2025
Revised: 17 March 2025
Accepted: 19 March 2025
Published: 25 March 2025

Citation: Luo, X.; Zhu, X. YOLO-SMUG: An Efficient and Lightweight Infrared Object Detection Model for Unmanned Aerial Vehicles. *Drones* **2025**, *9*, 245. https://doi.org/10.3390/drones9040245

1. Introduction

With the advancements in unmanned aerial vehicle (UAV) technology and infrared thermal imaging, UAVs equipped with infrared cameras have become increasingly lightweight and compact. Infrared-based object detection in UAV applications has shown great potential across various domains, providing advantages such as cost-effectiveness, operational flexibility, and improved detection performance in fields such as medical care [1], rescue operations [2], power systems [3], firefighting [4], agriculture [5], and national defense [6]. Based on these applications, researchers from various domains have extensively studied and improved UAV-based object detection.

For instance, in the medical field, Kotlinski M., et al. [7] proposed that the implementation of U-Space and UTM, combined with the application of infrared object detection technology, enables UAVs to identify and avoid obstacles in complex or low-visibility environments, thereby ensuring the safe delivery of medical supplies. In rescue operations, Polukhin A., et al. [8] analyzed the performance of various YOLOv5 deep learning model architectures and underscored the potential of integrating infrared detection technology

with UAVs to enhance rescue operations. The improved object detection capabilities can significantly aid in locating individuals in need of assistance. In the power system sector, Liao K.C. and Lu J.H. [9] proposed a method for detecting solar module faults in solar power farms using UAVs equipped with infrared and visible light cameras. This method enables real-time monitoring and analysis, enhancing maintenance efficiency and power generation. In the firefighting field, Ma Y., et al. [10] developed a visual algorithm utilizing infrared imaging for firefighting UAVs, which enhances fire detection capabilities through improved visual analysis. In agricultural applications, Messina G. and Modica G. [11] reviewed the state-of-the-art applications of UAV-based thermal imagery in precision agriculture. They highlighted its effectiveness in monitoring plant water stress, detecting diseases, estimating crop yields, and conducting plant phenotyping. In national defense, Christnacher F., et al. [12] explored the integration of optical and acoustical detection methods for UAVs, which enhance national defense capabilities through improved detection and tracking.

Currently, deep learning algorithms for UAV object detection can be divided into two types: one-stage and two-stage algorithms. Two-stage object detection models, such as R-CNN [13], Fast R-CNN [14], and Faster R-CNN [15], rely on region proposal mechanisms to identify objects; while these methods achieve high accuracy, they involve computationally intensive processes, making them unsuitable for UAV applications where real-time processing and lightweight deployment are essential. Their reliance on region proposals leads to higher latency, limiting their applicability in resource-constrained UAV systems. In contrast, one-stage models, including SSD [16], RetinaNet [17], and YOLO [18], eliminate the need for region proposals by directly predicting object locations and classifications from the input image. This architecture significantly reduces inference time, making one-stage models more suitable for UAV-based object detection. Additionally, these models incorporate multi-scale feature fusion, enhancing their ability to detect small objects—an essential capability for UAV applications where objects may vary in size and appear against complex backgrounds. Since YOLO serves as the foundation for our proposed model, it is particularly advantageous for UAV scenarios due to its single-shot detection framework, efficient convolutional architecture, and real-time processing capabilities. YOLO identifies objects by dividing the input image into a grid and predicting bounding boxes and class probabilities for each cell, allowing for fast and accurate detection even in dynamic UAV environments.

In drone-based infrared target detection, infrared images often exhibit complex backgrounds, substantial noise, limited color features, low resolution, and weak target edges and texture details. These factors collectively pose significant challenges for traditional detection algorithms, making it difficult to achieve high detection accuracy while maintaining a low miss rate. To improve accuracy and practicality, researchers have explored various methods tailored to UAV applications in diverse environments.

To address the challenges posed by complex backgrounds and high noise levels, Ma et al. [19] extracted deer targets to eliminate background interference and then employed a classification network (CA-Hybrid) based on Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and Channel Attention Mechanisms (CAMs) for accurate species identification of deer. Yang et al. [20] proposed a UAV infrared target detection framework based on Continuous Coupled Neural Networks (CCNNs), which simulates the processing mechanism of the human brain's visual cortex. This framework encodes the grayscale values of image pixels as neuronal firing frequencies and differentiates targets by analyzing the frequency variations among neural clusters, thereby enhancing detection performance in complex environments. Fang et al. [21] introduced a novel multi-scale U-Net architecture that models UAV target detection as a task of predicting residual images (including background, clutter, and noise). By learning the nonlinear mapping from input

images to residual images, this approach effectively mitigates the impact of severe infrared image noise on detection accuracy.

To address the challenges of limited color features and low resolution, Zhang et al. [22] introduced the Squeeze-and-Excitation (SE) module and enhanced the feature pyramid structure. By integrating multi-scale feature fusion, they expanded the network's receptive field and improved the recognition accuracy of small targets. Xu et al. [23] incorporated the latest Retentive Network Meets Transformer (RMT) technology into the backbone of YOLOv9s, enabling the extraction of both local and global features to enhance small target detection capabilities. Niu et al. [24] proposed an improved YOLOv5s-Seg model (FFDSM), which utilizes global average pooling and linear transformations to extract global information while adaptively adjusting channel weights. These enhancements strengthened feature representation, improving UAV-based infrared detection of forest fires. Liu et al. [25] developed their method based on YOLOv3 to address the performance degradation in small object detection caused by limited features. They optimized the Resblock in Darknet by connecting two ResNet units with identical width and height. Moreover, they enhanced the entire Darknet architecture by adding convolution operations in the early layers to enrich spatial information, which helped to expand the receptive field. These modifications were aimed at improving the detection of small targets.

To address the challenges of weak edge and texture information, Pan et al. [26] addressed the issue by adding noise to original low-quality images to generate paired high- and low-quality images. Using a Generative Adversarial Network (GAN), they learned image restoration and target feature enhancement, thereby improving UAV infrared small target detection performance. Zhang et al. [27] introduced the BSCA module to filter out instances of abnormal activation caused by occlusions and focus on pixels at the same spatial locations across different channels. This enhancement strengthened the model's ability to detect edge-blurred or overlapping targets. Wang et al. [28] introduced a technique that regulates the speed of a quadrotor; by maintaining the speed within a specific range identified through experimentation, this method reduces motion blur and enhances image clarity, thereby improving crack identification and addressing the issue of poor edge and texture visibility caused by motion blur in wind turbine blade inspections.

Nevertheless, processing infrared thermal images demands considerable computational resources, whereas drones have limited onboard computing capabilities. If the detection model is too large, deploying it on such mobile platforms becomes challenging. Therefore, this study proposes a lightweight UAV object detection method that focuses on balancing detection accuracy and reducing model complexity. The key contributions are presented here.

In the backbone part, a new network architecture based on ShuffleNetV2 is introduced. This architecture is enhanced by adding the MSDA mechanism. This allows the model to extract target features from infrared images more effectively, even in complex environments with cluttered backgrounds and varying object scales, thereby improving detection accuracy. As a result, the model can keep high detection accuracy while using fewer parameters and less computation. This makes it a good choice for environments with limited resources, especially for UAVs.

In the neck part, we use a new structure called C2f_UIB and the GhostConv module. These replace the old C2f structure and regular convolutional layers. This modification simplifies the model and reduces computational cost and parameter count. However, it still works well in finding and combining important features from different sizes; thus, it can better deal with the problems of lack of color characteristics, blurred edges, and low resolution of infrared images.

In the loss function part, the standard CIoU loss is substituted with the MPDIoU loss. This change makes object localization more accurate by better measuring the distance between predicted boxes and true boxes. It helps the model place objects more precisely and cuts down on both false positives and false negatives, at the same time reducing the detection error due to more noise in the infrared images.

2. Model and Training

2.1. YOLO-SMUG Model Network Structure

To effectively address the challenges associated with infrared imaging—such as low contrast, susceptibility to environmental variations, high noise levels, and low resolution—while simultaneously overcoming the computational, power processing, and data transmission constraints of UAV platforms, this study introduces YOLO-SMUG, a lightweight infrared object detection model based on YOLOv8 (small). The optimized architecture, illustrated in Figure 1, incorporates several key enhancements to improve detection accuracy and efficiency, making it well suited for deployment on resource-constrained UAV systems:

- Backbone Optimization: A novel backbone network based on ShuffleNetV2 is integrated with a Multi-Scale Dilated Attention (MSDA) mechanism to improve the model's ability to extract essential features from infrared environments with complex, cluttered backgrounds and variations in object scale and size. This modification significantly reduces the number of parameters and computational complexity while maintaining high detection accuracy, ensuring that the model remains suitable for efficient processing on UAVs with limited onboard computing resources.
- Neck Component Enhancement: The conventional C2f and convolutional layers are replaced with a lightweight inverted bottleneck structure, C2f_UIB, and GhostConv, effectively minimizing computational overhead and parameter size without compromising the model's feature extraction capabilities. This structural refinement enhances processing efficiency, enabling the UAV to better address issues such as insufficient color characteristics, blurred edges, and low resolution in infrared images.
- Improved Loss Function: The original loss function is substituted with MPDIoU, which enhances object localization accuracy by refining bounding box regression. This improvement is crucial for practical applications, as it ensures precise object identification and tracking in real-world scenarios, while also minimizing detection errors caused by the higher noise levels in infrared images.

These optimizations collectively enable the UAV to perform effective object detection using infrared imagery, even in the absence of high-performance embedded computing hardware. The model's lightweight nature ensures it can function effectively within the constraints of UAVs, utilizing efficient inference strategies to achieve processing with minimal computational demand.

2.2. Designing a Lightweight Feature Extraction Network with an Integrated Attention Mechanism

Infrared-image-based object detection plays a crucial role in UAV applications, where effective processing and high accuracy are essential despite constrained computational resources. However, the complexity and cluttered nature of infrared backgrounds, along with the varying scale and proportions of objects, pose significant challenges for robust detection. To address these challenges, this paper proposes a novel backbone network based on ShuffleNetV2, enhanced with a Multi-Scale Dilated Attention (MSDA) mechanism. This design significantly improves feature extraction efficiency while reducing the number of parameters and computational complexity. By maintaining high detection accuracy,

our approach ensures the model's suitability for UAV deployment, enabling robust and efficient infrared image analysis in real-world scenarios.

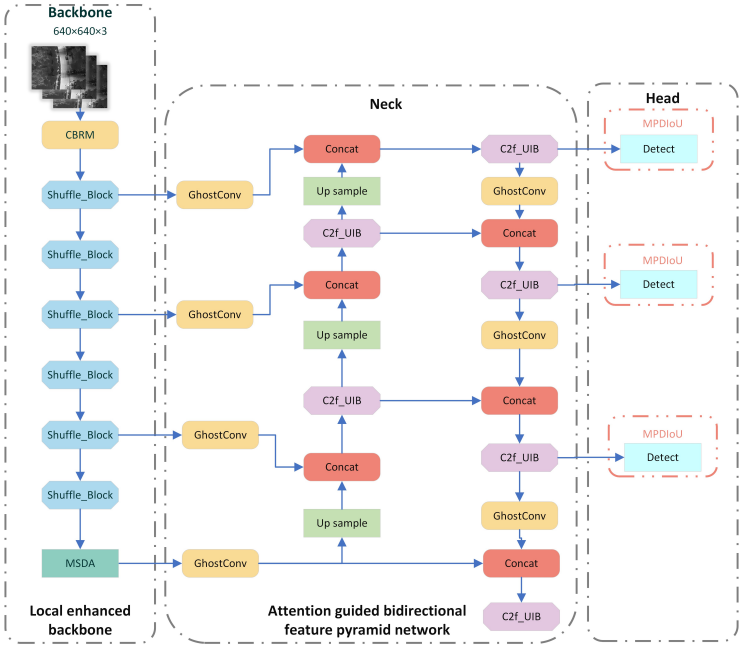


Figure 1. The structure of YOLO-SMUG, where Shuffle_Block utilizes channel shuffle and group convolution to enhance computational efficiency, MSDA integrates multi-scale dilated convolutions with attention mechanisms to improve feature extraction, GhostConv generates redundant feature maps through cheap operations to reduce computational cost while maintaining feature representation, and C2f_UIB, an enhanced Faster Implementation of CSP Bottleneck with 2 Convolutions Universal Inverted Bottleneck (C2f_UIB) module, is designed for infrared object detection in UAV systems.

2.2.1. ShuffleNetv2 Model

Ma et al. introduced the ShuffleNetv2 module [29] as an improvement on their previous ShuffleNetv1 model. As shown in Figure 2, the ShuffleNetv2 module is mainly composed of two parts: the basic unit and the down-sampling unit.

In the basic unit (Figure 2(1)), the input feature channels are divided into two groups. One group undergoes identity mapping, while the other passes through a sequence of operations: a 1×1 standard convolution, a 3×3 depthwise separable convolution, and activation layers. This structure facilitates efficient feature fusion across channels while minimizing computational cost. Uneven channel splitting prevents network fragmentation and improves efficiency, while maintaining a consistent channel width across layers reduces memory overhead and enhances communication between feature groups.

The down-sampling unit (Figure 2(2)) differs by omitting the channel splitting step. Instead, it applies a 3×3 separable depthwise convolution, a 1×1 basic convolution, and activation layers to the previously untouched branch. This operation halves the spatial dimensions while doubling the number of channels. After merging the two branches, a channel shuffle operation is performed to enhance information flow, increasing the network's width and improving feature extraction with minimal computational overhead.

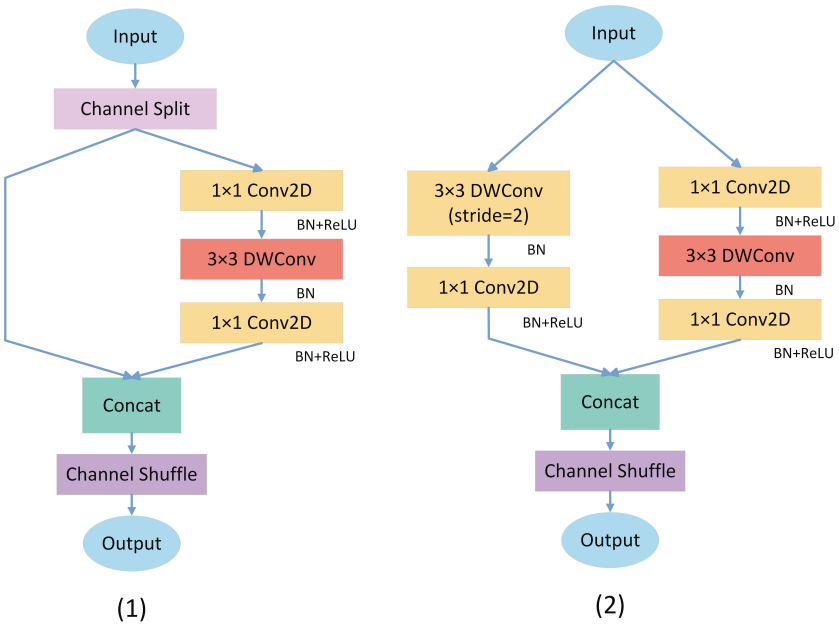


Figure 2. ShuffleNetv2 unit. (1) Basic unit. (2) Down-sampling unit.

2.2.2. Multi-Scale Dilated Attention Mechanism

Remote sensing infrared images often present complex and dynamic backgrounds, making accurate object extraction a challenging task. Moreover, these images typically feature objects with significant variation in scale and size, where the size of the receptive field plays a critical role in detection performance [30]. When the object size exceeds the receptive field, feature extraction becomes insufficient; in contrast, if the receptive field is too large, background noise can interfere with object recognition, leading to degraded performance. As a result, employing a fixed-size convolution kernel for object detection is less desirable.

To overcome these challenges, we propose the Multi-Scale Dilated Attention (MSDA) mechanism. This mechanism improves the network's ability to adapt to channel correlations and enhances the weighting of important channel features. Doing this makes the model better at obtaining important features and improves detection accuracy. MSDA cuts down on interference from busy backgrounds. This helps the detector focus on useful features and keeps a good balance between computational complexity and receptive field size.

In MSDA, the feature map's channels are split into n distinct heads, each applying multi-scale dilated convolutions with varying dilation rates. This is called Scaled Dilated Attention (SWDA). This approach enables the model to capture image features across multiple scales. It improves the model's ability to detect objects of different sizes. The structure of MSDA is illustrated in Figure 3.

As shown, MSDA first divides the feature map's channels into several groups. Every group has its own attention head. Different dilation rates, like 1, 3, and 5, are used for each head. This lets the model capture information from different areas and sizes. It helps the model focus on different parts and obtain more useful features. Then, self-attention is used in each head to give more importance to the most important features. This fixes the problem with old attention methods that only look at nearby features and miss important far-away information.The SWDA operation is applied to each attention head, as given by the following equation:

$$h_a = \mathrm{SWDA}(Q_a, K_a, V_a, r_a), \quad 1 \leq a \leq n \tag{1}$$

where Q_a, K_a, V_a represent the query, key, and value for each attention head, respectively, and r_a is the dilation rate corresponding to each head (e.g., $r_1 = 1, r_2 = 3$). The output for each attention head is denoted by h_a.

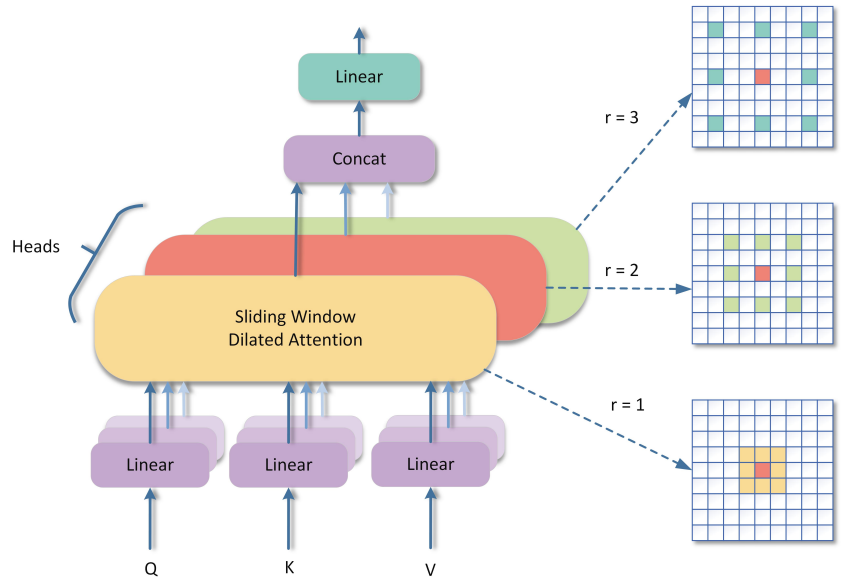


Figure 3. Illustration of MSDA.

Before splitting, the SWDA operation is performed on the point (m, n) within the initial feature map and is formulated as follows:

$$X_{mn} = \text{Attention}(Q_{mn}, K_q, V_q) = \text{Softmax}\left(\frac{Q_{mn}K_r^T}{\sqrt{d_K}}\right) \times V_q, \quad 1 \le m \le W,\ 1 \le n \le H \quad (2)$$

Here, W and H are the width and height of the feature map, respectively, and K_q and V_q are the keys and values selected from the feature maps K and V. In this equation, the Softmax operation is computed over the spatial dimension, meaning that, for each query at position (m, n), the attention weights are normalized across all key positions (m', n'). This ensures that the attention mechanism focuses on the most relevant spatial locations within the feature map.

For each query at position (m, n), the keys and values for self-attention are selected from the coordinates (m', n'), where:

$$\{(m', n') \,|\, m' = r \times vs. + m,\ n' = vs. \times q + n\}, \quad -\frac{w}{2} \le r, vs. \le \frac{w}{2},\ r, vs. \in \mathbb{Z} \quad (3)$$

Here, w represents the dimensions of the sliding window used for convolution. This formulation describes a non-standard dilation mechanism that differs from traditional dilated convolutions in CNNs. In standard dilated convolutions, the dilation rate r is applied uniformly to the kernel, expanding the receptive field by skipping fixed intervals in the input feature map. However, in this method, the dilation mechanism is adaptive and query-dependent. Specifically, for each query at position (m, n), the keys and values are selected from dynamically calculated coordinates (m', n'), which are determined by the parameters r, $vs.$, and q. This allows the model to capture more flexible and context-aware spatial relationships, as the dilation pattern varies depending on the query position and the learned parameters. Unlike standard dilated convolutions, which apply a fixed dilation pattern across the entire feature map, this approach enables the model to adaptively adjust the receptive field based on the local context, potentially improving performance on tasks requiring fine-grained spatial understanding.

MSDA efficiently combines them through a linear layer, following the extraction of multi-scale features through the self-attention mechanism. This process allows the model to capture both local details and the global structure of the image, resulting in a more comprehensive understanding of the infrared image content [31].

2.2.3. Improved Feature Extraction Network

To summarize, the backbone of our network was specifically designed to address the unique challenges in remote sensing object detection, especially for complex, cluttered backgrounds and significant variations in object scale and size. The redesigned backbone starts with a Convolutional–BatchNorm–ReLU–MaxPooling (CBRM) module, which efficiently extracts low-level features while maintaining computational efficiency. This is followed by six Shuffle_Block modules, which strike an effective balance between accuracy and efficiency. These Shuffle_Block modules enhance feature extraction by utilizing lightweight operations such as channel splitting, shuffling, and merging. These operations improve feature representation without a significant increase in computational cost, making the design especially suitable for remote sensing tasks where large input images need to be processed efficiently to maintain excellent performance.

To make the model better at detecting objects of different sizes, especially small or faraway ones, we added a Multi-Scale Dilated Attention mechanism at the backbone's final layer. MSDA uses multi-scale dilated convolutions. This helps the network capture information from different areas. It makes sure small objects are not missed and keeps the big picture for large objects. By changing the receptive fields, MSDA helps the network handle the many different object sizes in remote sensing images.

Moreover, the MSDA mechanism employs an attention mechanism that assigns higher weights to critical features while suppressing background noise. This improves the network's focus on the most discriminative regions of the image. MSDA makes the important features stronger. This helps the network tell objects apart from complex backgrounds and cuts down the effect of unimportant or confusing details in crowded scenes. These improvements make the feature extraction better and more reliable, resulting in improved detection accuracy and better generalization across a wide range of remote sensing datasets.

2.3. Designing a Enhanced Feature Fusion Network

2.3.1. Introducing the Lightweight Ghost Convolution Module

In the YOLOv8 structure, Cross-Stage Partial (CSP) networks are used in the feature fusion module [32]; while the CSP module helps reduce the number of parameters to some extent, when faced with issues such as low resolution and blurred edges in infrared images, it cannot effectively perform object detection, and the overall network size remains too large for efficient deployment in UAV object detection applications.

Regarding network compression and pruning, many studies have shown that feature maps generated by popular CNN architectures often contain significant redundancy, with some features being highly similar [33]. To address this, GhostNet proposes that these redundant features are necessary for high detection performance in some networks [34]. Based on this idea, GhostNet introduces the GhostConv module, which aims to reduce feature redundancy while maintaining performance [35]. This module, depicted in Figure 4, provides a more efficient way of processing features without sacrificing detection accuracy.

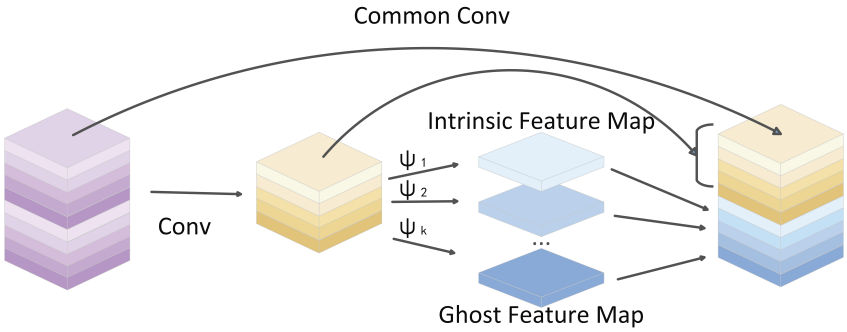


Figure 4. Illustration of Ghost Convolution mechanism.

We donate the input feature map by $X \in \mathbb{R}^{h \times w \times c}$, where h and w denote the height and width, and c represents the number of input channels. The output feature map is given by $Y' = X \cdot f$, where $f \in \mathbb{R}^{c \times k \times k \times m}$ represents the convolution kernel, and $Y' \in \mathbb{R}^{m \times h' \times w'}$. Here, m represents the number of output channels, and h' and w' are the dimensions of the output feature map. The generation of the ghost feature map Y' is defined by Equation (4):

$$y_{ij} = \phi_{ij}(y'_i), \quad \forall i = 1, \ldots, m, \; j = 1, \ldots, s \tag{4}$$

In Equation (4), m represents the number of channels in Y', y'_i denotes the i-th channel, and j refers to the j-th linear transformation applied to y'_i.

Each feature map y'_i undergoes $s-1$ linear transformations, followed by an identity transformation, resulting in the final output, Y, with $n = m \cdot s$ channels. Therefore, the total number of linear transformations applied to the feature maps is $m \cdot (s-1) = \frac{n}{s} \cdot (s-1)$.

Assuming a kernel size of $d \times d$, the theoretical computational speed ratio r_s and parameter compression ratio r_c for the Ghost module are given by:

$$\begin{aligned} r_s &= \frac{n \cdot h' \cdot w' \cdot c \cdot k \cdot k}{\frac{n}{s} \cdot h' \cdot w' \cdot c \cdot k \cdot k + (s-1)\frac{n}{s} \cdot h' \cdot w' \cdot d \cdot d} \\ &= \frac{c \cdot k \cdot k}{\frac{1}{s} \cdot c \cdot k \cdot k + \frac{(s-1)}{s} \cdot d \cdot d} \approx \frac{s \cdot c}{s + c - 1} \approx s \\ r_c &= \frac{n \cdot c \cdot k \cdot k}{\frac{n}{s} \cdot c \cdot k \cdot k + (s-1)\frac{n}{s} \cdot d \cdot d} \\ &\approx \frac{s \cdot c}{s + c - 1} \approx s \end{aligned} \tag{5}$$

In the derivation of r_s, we assume a kernel size of $d \times d$, specifically referring to `pointwise convolution` (with a kernel size of 1×1) rather than `depthwise convolution` (with a kernel size of $d \times d$). Pointwise convolution involves only linear transformations along the channel dimension, without spatial convolution, which significantly reduces computational cost compared to depthwise convolution. This assumption is critical for the derivation of r_s. The approximation $r_s \approx s$ is based on the assumption that $s \gg c$ or that s dominates the denominator. In deriving r_s, we assume that $s \gg c$ or that s dominates the denominator. This assumption is reasonable under the following conditions:

1. When $\boldsymbol{s}$ is much larger than $\boldsymbol{c}$: That is, when the number of linear transformations $\boldsymbol{s}$ is much greater than the number of input channels $\boldsymbol{c}$. In this case, the $\boldsymbol{s}$ term in the denominator dominates; hence, $\boldsymbol{r_s \approx s}$.
2. When $\boldsymbol{c}$ is relatively small: That is, when the number of input channels $\boldsymbol{c}$ is small, and $\boldsymbol{s}$ is relatively large. At this point, the $\boldsymbol{c}$ term in the denominator has a minimal impact on the overall result, and the approximation $\boldsymbol{r_s \approx s}$ still holds.

3. In lightweight networks: Lightweight networks are typically designed with a larger ***s*** and a smaller ***c*** to meet the requirements for computational efficiency. Therefore, this assumption has practical significance in the design of lightweight networks.

However, when c is very large (for example, in deep networks with a high number of channels), this assumption may no longer hold, as the c term in the denominator can significantly affect the outcome. In such cases, it is necessary to re-evaluate the calculation method for r_s, or to verify the rationality of the approximation through experiments.

From these equations, it is clear that Ghost Convolution enhances computational speed and reduces parameters by a factor of s without compromising the network's ability to extract features effectively. The efficiency of Ghost Convolution makes it particularly suitable for resource-constrained applications, such as mobile and embedded devices.

2.3.2. Introducing of the Lightweight Module C2f_UIB

Transformer models often require a large size of parameters during the process of training, which can make them less efficient for operation on mobile devices. To address the challenge while maintaining detection accuracy, this work presents an enhanced Faster Implementation of CSP Bottleneck with 2 Convolutions Universal Inverted Bottleneck (C2f_UIB) module. This module substitutes the initial C2f unit in the neck part of the baseline YOLOv8s model. The designs of the C2f and C2f_UIB modules are depicted in Figure 5.

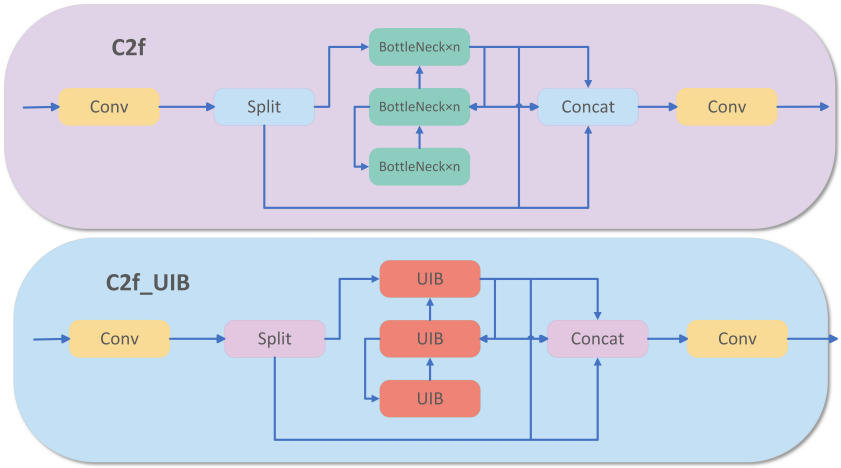


Figure 5. C2f and C2f_UIB module structure diagrams.

The Universal Inverted Bottleneck (UIB) [36], initially presented in MobileNetV4, is part of a lightweight end-to-end network developed by Google for mobile device applications; it plays a crucial role in reducing model complexity. As shown in Figure 6, the UIB module integrates two optional depthwise (DW) convolutions within the inverted bottleneck block (IB), greatly reducing the number of parameters and computational cost. One depthwise convolution is positioned before the expansion layer, and the other is located between the expansion and projection layers. This architecture combines features in MobileNetV2's inverted bottleneck, ConvNext blocks, and Vision Transformer (ViT) feed-forward networks. Additionally, the extra depthwise (ExtraDW) structure is introduced to further reduce the model's computational load.

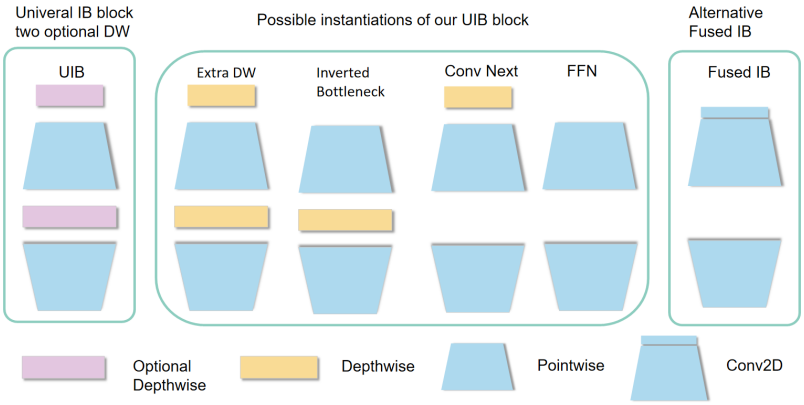


Figure 6. Universal Inverted Bottleneck(UIB) module.

2.3.3. Improved Feature Fusion Networks

In summary, we propose a method to integrate two lightweight modules—C2f_UIB and GhostConv—into the neck architecture of YOLOv8 in our study. GhostConv is particularly beneficial for infrared image processing due to its ability to extract rich features with minimal computational cost. Unlike standard convolution, which directly computes all feature maps [37], GhostConv first extracts primary feature maps and then generates additional ghost feature maps through lightweight transformations. This mechanism effectively enhances feature diversity without significantly increasing computational complexity. In infrared images, where texture and structural information are important due to the lack of color cues, GhostConv helps to recover fine-grained details by reinforcing feature redundancy and preserving edge structures [38]. The generated ghost feature maps capture complementary spatial details, helping to counteract the blurred edges and low resolution often found in infrared imagery.

On the other hand, C2f_UIB further improves feature aggregation in infrared object detection [39]. Traditional C2f structures rely on cross-stage feature fusion but may introduce redundant information, which can be problematic for low-resolution infrared images [40]. C2f_UIB optimizes this process by employing a Unidirectional Information Broadcast (UIB) mechanism, which selectively propagates high-level semantic features while maintaining critical low-level spatial details. This structured information flow is particularly useful for infrared images, as it enhances object boundaries and improves detection robustness despite poor edge contrast. Furthermore, C2f_UIB utilizes lightweight convolutions [41], reducing computational cost while ensuring that essential geometric and texture details are retained—a crucial aspect for detecting small or low-contrast objects in infrared imagery.

By integrating GhostConv and C2f_UIB into the YOLOv8 neck, our model achieves a better trade-off between feature richness and computational efficiency, making it well suited for real-time infrared object detection. GhostConv compensates for missing color features by enhancing spatial feature diversity, while C2f_UIB improves feature fusion and boundary sharpness, addressing key challenges of infrared image processing. Experimental results demonstrate that our modifications lead to improved detection performance on infrared datasets, particularly in terms of edge preservation, small object detection, and robustness to low-resolution inputs.

2.4. Designing of MPDIoU-Based Loss Function

In initial YOLOv8s, the Complete Intersection over Union (CIoU) loss function calculates regression errors based on three key factors: overlap area, center point distance, and aspect ratio of the predicted bounding boxes [42]; while this formulation improves

localization precision, it also introduces an inherent dependency on aspect ratio alignment, which can become problematic in cases where the predicted and ground-truth bounding boxes differ significantly in scale. This issue is further exacerbated when dealing with small objects or noisy infrared images, where aspect ratio variations may not be reliable indicators of object alignment. Consequently, CIoU may over-penalize predictions that have an appropriate location but deviate in shape, leading to suboptimal bounding box regression [43].

To mitigate these issues, this paper introduces the MPDIoU loss function, which replaces CIoU by incorporating minimum point distance and intersection-over-concurrency ratios.

The MPDIoU metric consists of three crucial components: the consideration of both overlapping and non-overlapping regions, the calculation of centroid distances, and the assessment of deviations in width and height. By minimizing the distance between predicted and ground-truth bounding boxes, the approach simplifies the computation of Intersection over Union (IoU). The formulas for the MPDIoU loss function are presented in Equations (6) and (7).

$$L_{\text{MPDIoU}} = 1 - \text{MPDIoU} = 1 - \left(\frac{A \cap B}{A \cup B} - \frac{d_1^2}{w^2 + h^2} - \frac{d_2^2}{w^2 + h^2} \right) \tag{6}$$

$$|C| = \left(\max(x_2^{gt}, x_2^{prd}) - \min(x_1^{gt}, x_1^{prd}) \right) \times \left(\max(y_2^{gt}, y_2^{prd}) - \min(y_1^{gt}, y_1^{prd}) \right) \tag{7}$$

In Equation (6), $A \cap B$ represents the intersection area between the ground truth box (Box A, the yellow box in Figure 7) and the predicted box (Box B, the red box in Figure 7), while $A \cup B$ represents their union area. The terms d_1 and d_2 denote the deviations in width and height, respectively, between the predicted box and the ground truth box. These deviations are calculated as the differences in their center points along the horizontal and vertical axes. By minimizing these deviations, the MPDIoU loss function aims to improve the alignment between the predicted and ground truth boxes.

In Equation (7), $|C|$ represents the coverage area of B_{gt} and B_{prd}, which corresponds to the area of the smallest enclosing rectangle. This is illustrated in Figure 7 with a schematic diagram depicting the positional relationship between each parameter. Here, Box A represents the ground truth box, and Box B represents the predicted box. In addition, x_1^{gt} and x_2^{gt} represent the x-coordinates of the top-left and bottom-right corners of the ground truth box (Box A), while y_1^{gt} and y_2^{gt} represent the corresponding y-coordinates. Similarly, x_1^{prd} and x_2^{prd} denote the x-coordinates of the top-left and bottom-right corners of the predicted box (Box B), and y_1^{prd} and y_2^{prd} denote the corresponding y-coordinates. These coordinates are typically absolute, representing specific pixel positions in the image, but they may also be normalized, i.e., relative coordinates scaled to the image dimensions, usually within the range [0, 1].

Consequently, compared to the CIoU loss function, the MPDIoU loss function reduces computational complexity by directly measuring the pointwise distance between predicted and ground-truth bounding boxes, eliminating the need for complex geometric calculations involving aspect ratios or angles. This reduction in computational complexity enhances efficiency, making it particularly beneficial for effective infrared object detection.

More importantly, MPDIoU is inherently more robust to high noise levels in infrared images. Traditional IoU-based loss functions can be sensitive to shape distortions and inaccurate boundary predictions caused by noise, but MPDIoU focuses on minimizing the discrepancy between key points of the bounding boxes, reducing the impact of edge distortions, blurred object contours, and background clutter [44]. By refining the object localization process with a direct and noise-resistant distance metric, MPDIoU signifi-

cantly improves positioning accuracy, ensuring better detection performance in challenging infrared environments.

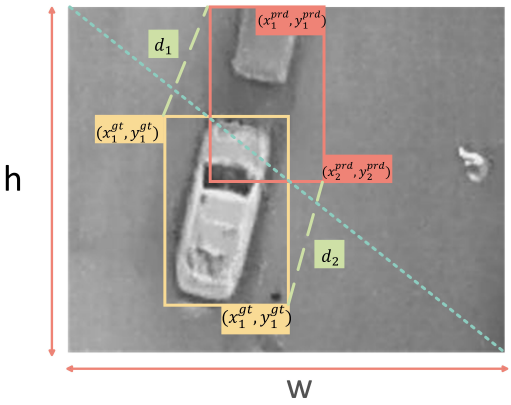


Figure 7. MPDIoU position relationships.

3. Model Training and Evaluation Metrics

3.1. Datasets

This study focuses on model optimization and design using the HIT-UAV dataset [45]. Released publicly by Nature Research in April 2023, this dataset is the first high-altitude UAV infrared dataset, as shown in Figure 8.

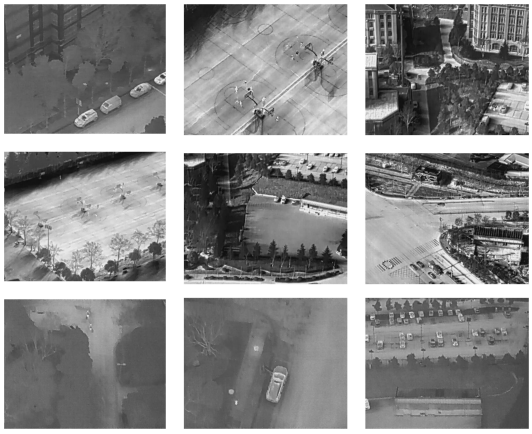


Figure 8. Partial image presentation of the HIT-UAV dataset.

This dataset is a specialized dataset for UAV-based object detection, which includes infrared thermal images captured from diverse environments such as schools, parking lots, streets, and playgrounds. In addition, the dataset is divided into five classes: Person, Car, Bicycle, Other Vehicle, and Do not Care [46,47].

The updated dataset has 2898 images, each with a resolution of 640 × 512 pixels. In this study, the dataset is divided into three sections following a 7:2:1 ratio. In total, there are 2008 images allocated for training, 571 for testing, and 287 for validation.

3.2. Assessment Indicators

This paper evaluates model performance using precision (P), recall (R), F1-score, average precision (AP), mean average precision (mAP), parameter count, and GFLOPs. Model size and computational complexity are quantified by parameters and GFLOPs, respectively. The evaluation metrics are defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

where TP is the number of true positives, and FP is the number of false positives. Precision measures the ratio of correctly predicted positives to all predicted positives.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

where FN is the number of false negatives. Recall measures the ratio of correctly predicted positives to all actual positives.

$$\text{F1-score} = \frac{2 \times P \times R}{P + R} = \frac{2TP}{2TP + FP + FN} \tag{10}$$

The F1-score is the harmonic mean of precision and recall, providing a balanced measure of model accuracy.

$$AP = \int_0^1 P(r), dr \tag{11}$$

Average precision (AP) is the area under the precision–recall curve, representing the model's performance across different recall levels.

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{12}$$

Mean average precision (mAP) is the average of AP values across multiple categories, indicating overall detection performance.

3.3. Experimental Platform

To evaluate the effectiveness of the YOLO-SMUG model, we designed both a comparative test and an ablation experiment. The hardware platform specifications used during these experiments are provided in Table 1.

Table 1. Experimental platform information.

Name	Related Configurations
CPU	12 vCPU Intel(R) Xeon(R) Silver 4214R
GPU	RTX 3080 Ti (12 GB)
RAM	30 GB
Language	Python 3.8 (Ubuntu 20.04)
Framework	PyTorch 1.11.0
CUDA Version	CUDA 11.3
epoch	280

3.4. Experimental Results

3.4.1. Comparative Analysis of the Incorporation of Different Attention Mechanisms

To investigate the rationale for introducing the MSDA mechanism, other representative attention mechanism networks, including SE, GAM, iRMB, ECA, and CA, were selected for experimental comparison and tested on the test set. The results of the comparative experiments are listed as Table 2.

Table 2. Performance metrics for different models.

Model	P	R	mAP_{50}	$mAP_{50–95}$	GFLOPS	Parameters/10^6
*	**0.837**	**0.774**	**0.818**	**0.525**	**12.1**	**3.165578**
*+SE	0.807	0.756	0.803	0.527	12.3	3.987487
*+GAM	0.810	0.734	0.801	0.525	13.7	5.628703
*+iRMB	0.811	0.726	0.803	0.512	20.1	4.254495
*+ECA	0.881	0.734	0.813	0.523	12.3	3.987490
*+CA	0.801	0.719	0.810	0.501	13.3	4.037180
***+MSDA**	**0.837**	**0.774**	**0.825**	**0.546**	**12.2**	**3.954719**

The * denotes models that include ShuffleNetV2, GhostConv, and C2f_UIB as baseline components.

From Table 2 and the visualization of some of their key indicators Figure 9, it can be clearly observed that the MSDA model demonstrates the best overall performance, with a 0.8% growth in mAP_{50} and a 4.00% increase in $mAP_{50–95}$, while only increasing GFLOPS by 0.83% and parameters by 24.93%, making it the most balanced improvement. The ECA model improves precision by 5.26%, but at the cost of a 5.17% drop in recall, making it suitable for scenarios where reducing false positives is critical. The SE model shows a slight 0.38% increase in $mAP_{50–95}$, but with a 1.8% decrease in mAP_{50} and minor drops in both precision and recall. On the other hand, GAM, iRMB, and CA significantly increase computational cost (+66.12% GFLOPS for iRMB, +78.02% parameters for GAM) while failing to improve or even reducing detection performance (mAP_{50} drops by -2.08% for GAM, mAP_{50} drops by -5.16% for iRMB and mAP_{50} drops by -0.98% for CA). Overall, MSDA is the best choice, followed by ECA, while GAM, iRMB, and CA are not recommended due to their high computational cost with little to no performance gain. So we choose MSDA as our best attention mechanism.

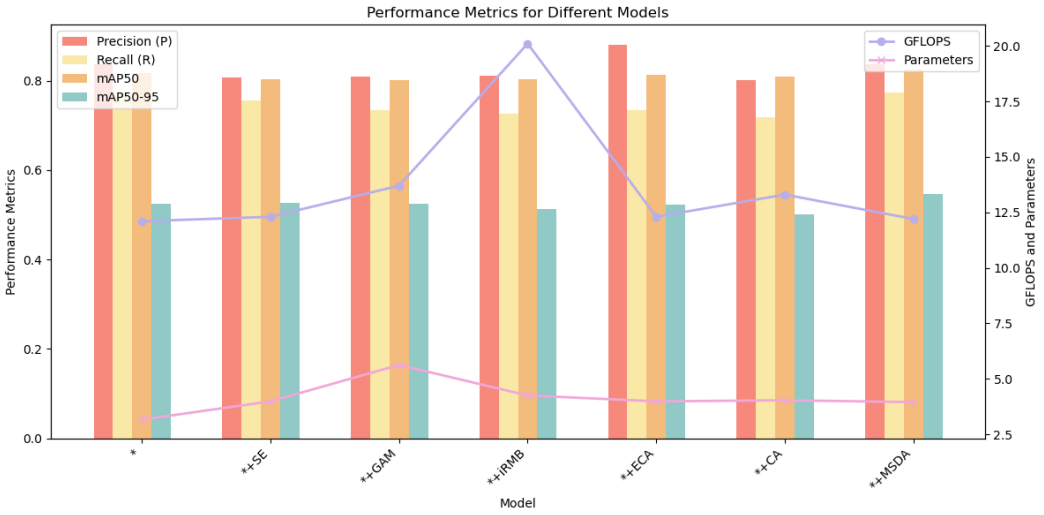


Figure 9. Normalization analysis of different models. The * denotes models that include ShuffleNetV2, GhostConv, and C2f_UIB as baseline components.

3.4.2. Comparative Analysis of the Incorporation of Varied Loss Function

In order to verify the efficacy of the proposed loss function, our study compared its impact on model performance against various alternatives, after optimizing the backbone and neck networks. As shown in Table 3 and Figure 10, introducing the WIOU loss function led to a significant 2.4% improvement in precision (P) (from 0.837 to 0.861) and a 0.5% increase in $mAP_{50–95}$ (from 0.546 to 0.551), but a slight decrease of 1.2% in recall (R) (from 0.774 to 0.762). This suggests that the WIOU loss function primarily enhances localization accuracy rather than global detection capability. FOCAL-GIOU improved precision to

0.899 (+7.4%), but it caused a significant drop in recall by 3.0% (from 0.774 to 0.751) and a 3.5% decrease in $mAP_{50–95}$ (from 0.546 to 0.527). This shows its weakness in focusing too much on hard samples. In contrast, MPD kept precision stable at 0.853 (+1.9%) while significantly improving recall to 0.778 (+0.5%) and mAP_50 to 0.84 (+1.8%), delivering the best overall performance. This suggests that the multi-scale feature fusion mechanism in MPD balances precision and recall better. Importantly, all these improvements were achieved without adding extra computational cost, confirming the lightweight nature of the loss function design. This experiment provides a quantitative basis for adapting loss functions in complex scenarios.

Table 3. Comparison of different loss functions.

Model	P	R	mAP_{50}	$mAP_{50–95}$
Shuffle+MSDA+UIB+GhostConv(*)	0.837	0.774	0.825	0.546
*+WIOU	0.861	0.762	0.833	0.551
*+FOCAL-GIOU	0.899	0.751	0.817	0.527
*+InnerMPD	0.858	0.765	0.827	0.534
*+MPD	0.853	0.778	0.84	0.538

[1] The * denotes models that include ShuffleNetV2, MSDA, GhostConv, and C2f_UIB as baseline components.

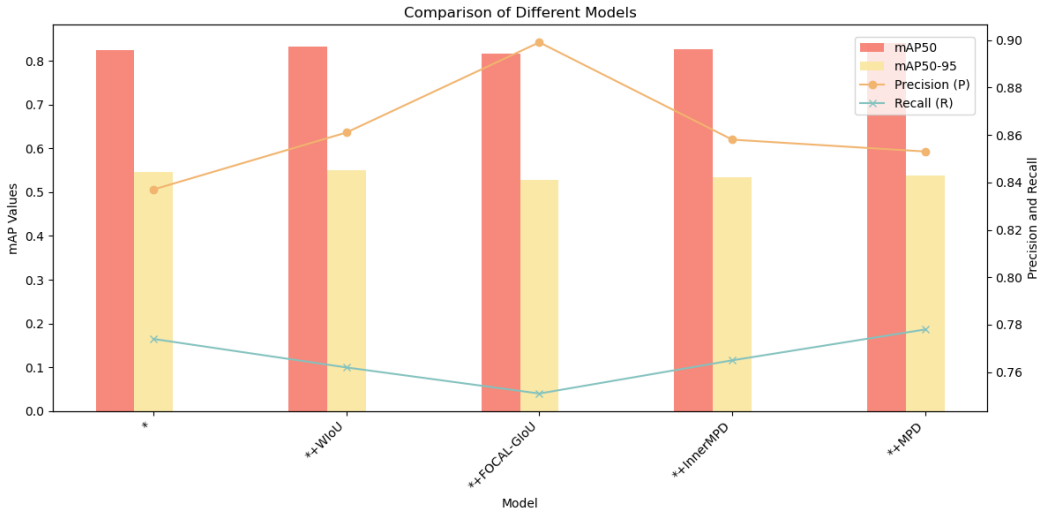


Figure 10. Normalization analysis of different loss functions. The * denotes models that include ShuffleNetV2, MSDA, GhostConv, and C2f_UIB as baseline components.

3.4.3. Ablation Experiment

To evaluate the effectiveness of each module in YOLO-SMUG, this study uses YOLOv8S as the baseline model and performs ablation experiments on ShufflenetV2, MSDA, C2f_UIB, GhostConv, and MPDIoU. The experiments were conducted on the HIT-UAV dataset, and the results are presented in Table 4. Each change helps improve detection and makes it easier to use on mobile devices.

Firstly, by incorporating the ShufflenetV2 module, the model's parameter count is reduced from 11.1 million to 5.7 million, a reduction of 48.53%, and the computational cost (GFLOPs) decreases from 28.4 to 15.9, a 44.01% reduction. Despite the significant decrease in both parameters and computation, mAP_50 only slightly decreases from 0.811 to 0.810, a drop of 0.12%. This result suggests that the lightweight structure of ShufflenetV2 only slightly affects the model's feature extraction efficiency, but significantly enhances its adaptability for deployment in resource-constrained mobile environments.

Next, by adding the MSDA module on top of ShufflenetV2, mAP_50 increases from 0.810 to 0.813, an improvement of 0.37%, while the F1-score rises by 2.60%. Despite a

3.31% increase in the parameter count and a 0.63% rise in the computational load, the MSDA module significantly enhances detection performance, particularly in its ability to strengthen the network's learning of shallow features, which is beneficial for detecting small objects. This demonstrates its crucial role in multi-scale information fusion and feature enhancement.

Adding the C2f_UIB module cuts parameter size by 46.46% and reduces GFLOPS by 20.63%. Meanwhile, mAP_50 goes up by 0.49%. This shows that the C2f_UIB module makes the network better by using less computation while keeping accuracy high.

Subsequently, the GhostConv module makes the model even better. The mAP_50 goes up from 0.817 to 0.825, which is a 0.98% increase. The amount of work the model has to do also goes down to 12.2 GFLOPs, a 3.94% reduction. The F1-score also improves by 2.53%. This shows that GhostConv helps the model work better and uses less computation.

In the end, after improving the loss function, the model's accuracy goes up to 0.84. This is 3.58% better than the starting model. The F1-score also rises from 0.77 to 0.82, which is a 6.49% increase. This means the model performs better in balancing precision and recall.

To sum up, the optimized YOLO-SMUG model is much better than the starting model. Accuracy goes up by 3.58%, and the F1-score improves by 6.49%. It also uses 57.04% less computation and has 64.38% fewer parameters. This means the model is faster and lighter, making it easier to use on mobile devices. It uses less memory and computing power while getting better at finding features. The model works really well on UAV infrared datasets, so it is perfect for drone-based tasks where resources are limited.

Table 4. Ablation experiment result.

Experimental ID	ShufflenetV2	MSDA	C2f_UIB	GhostConv	MPDIoU
1					
2	✓				
3	✓	✓			
4	✓	✓	✓		
5	✓	✓	✓	✓	
6	✓	✓	✓	✓	✓

Experimental ID	P	R	mAP_{50}	GFLOPS	Parameters/10^6	F1
1	0.820	0.784	0.811	28.4	11.127519	81
2	0.866	0.749	0.810	15.9	5.755798	77
3	0.877	0.751	0.813	16.0	5.945807	79
4	0.816	0.769	0.817	12.7	3.185590	79
5	0.837	0.774	0.825	12.2	3.954719	81
6	0.853	0.778	0.840	12.2	3.954719	82

3.4.4. Model Comparison Experiment

In order to showcase the advantages of the proposed algorithm, our study performed a comparative analysis between our improved model (YOLO-SMUG) and several leading state-of-the-art object detection algorithms, including the two-stage Faster R-CNN and multiple versions of the YOLO series (YOLOv5s, YOLOv5n, YOLOv6s, YOLOv6n, YOLOv8s, etc.) under the same experimental conditions. The statistical results are presented in Table 5.

As shown in Figure 11, the comparison reveals that Faster R-CNN falls behind the YOLO series in terms of model size, computational cost, and detection accuracy. YOLOv5n,

YOLOv6n, and YOLOv8n are good at saving computing power and having fewer parameters, but they do not perform well in pulling out features, so they do not detect things very well. On the other hand, YOLOv5s, YOLOv6s, and YOLOv8s are more accurate, but they struggle with finding small objects, which is a big deal for drone-based detection. Furthermore, they are too big and need too much computing power, so they are not good for using on drones or other devices that do not have a lot of resources.

Due to the limited computational resources of UAVs and the large size of detection models, UAV-based infrared object detection faces challenges in efficient processing. Our proposed model (YOLO-SMUG) outperforms the aforementioned networks by effectively balancing detection accuracy and efficiency. With only 3.9 M parameters and 12.2 GFLOPS, it remains lightweight while achieving the highest mAP_50 (84.0) and F1-score (82) among all compared models. This demonstrates that YOLO-SMUG enhances feature extraction capabilities without significantly increasing model size or computational complexity. These improvements make it a promising choice for UAV-based infrared object detection, where both performance and efficiency are crucial.

Table 5. Comparison of different models.

Model	Parameters/M	Size	GFLOPS	mAP_50	F1
Faster R-CNN	41.2	116.7	156.3	70.21	73
YOLOv5n	2.5	6.6	7.1	75.0	71
YOLOv5s	9.1	18.5	23.8	80.3	75
YOLOv6n	4.2	8.7	11.8	73.5	70
YOLOv6s	16.2	32.9	44.0	75.6	71
YOLOv8n	3.0	6.7	8.1	77.1	77
YOLOv8s	11.1	22.5	28.4	81.1	81
Ours	3.9	8.2	12.2	84.0	82

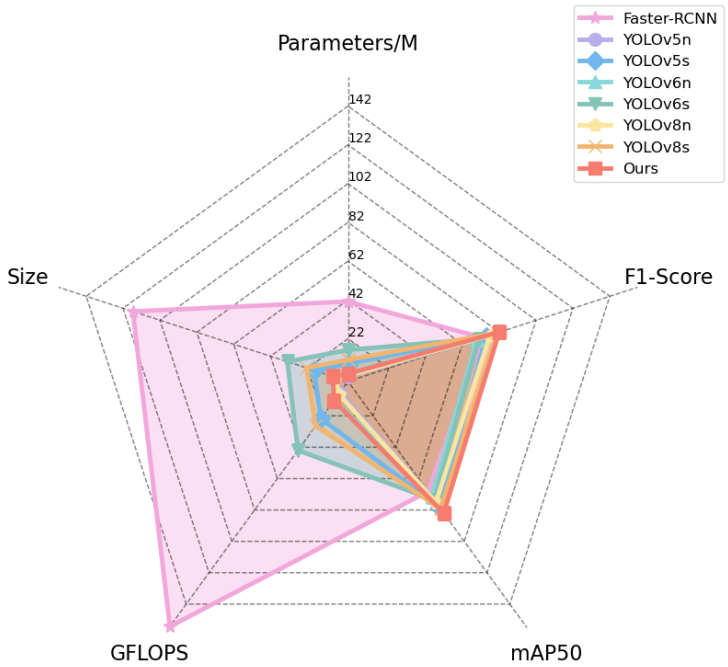
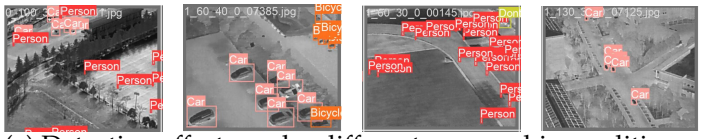


Figure 11. Visualization of performance of different models.

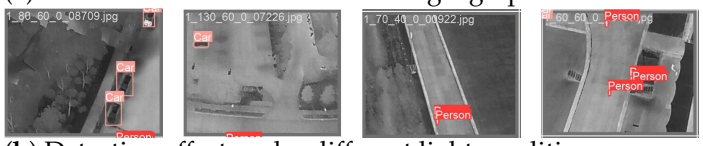
3.4.5. Visual Analysis

Drones encounter numerous challenges in complex environments, including variations in location, lighting conditions, flight altitude, and occlusions. To address these issues,

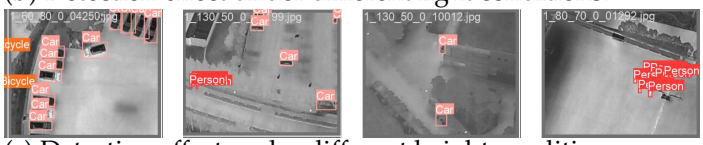
the improved model was evaluated across diverse scenarios, with the results presented in Figure 12. The findings demonstrate that the model effectively detects small and visually similar objects in drone-captured infrared images, highlighting its flexibility and reliability.



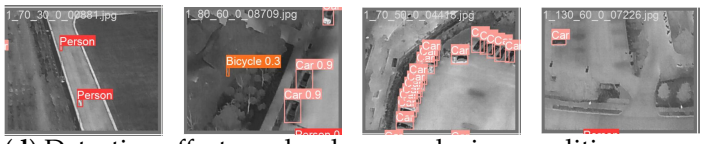
(**a**) Detection effects under different geographic conditions.



(**b**) Detection effect under different light conditions.



(**c**) Detection effect under different height conditions.

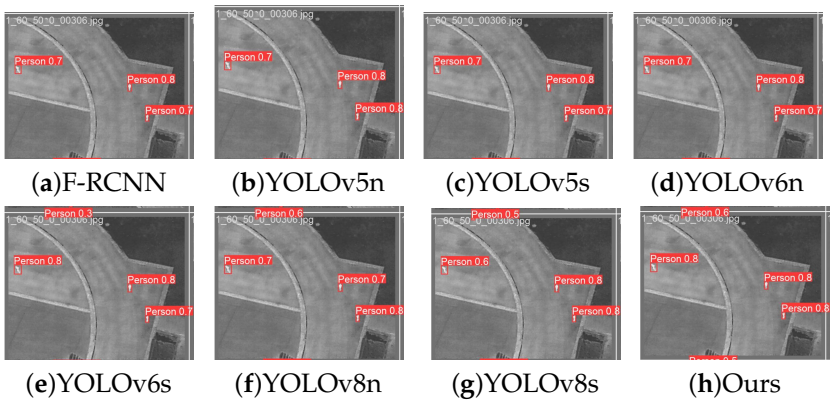


(**d**) Detection effects under dense occlusion conditions.

Figure 12. Some examples of infrared small object detection results across varied conditions.

This study evaluates the performance of drone-based infrared imaging technology in detecting small objects under various environmental conditions. The proposed algorithm demonstrates strong adaptability across diverse geographical settings, accurately identifying small objects in infrared imagery captured in scenarios such as streets and parking lots. Comparisons under different lighting conditions reveal that, while illumination changes influence detection outcomes, the algorithm remains robust due to the unique properties of thermal imaging, which are less affected by visible light variations. Additionally, an analysis of object height shows that higher altitudes increase occlusion and detection difficulty, yet the algorithm maintains high accuracy in processing infrared thermal images. Under dense occlusion, the algorithm successfully detects most obscured objects based on their heat signatures, though some remain undetected, suggesting room for improvement in handling severe occlusion. In summary, while the drone-based infrared small object detection algorithm performs effectively in challenging conditions, further refinements are essential in enhancing its practical applicability in specific scenarios.

Furthermore, Figure 13 presents the comparative small object detection results in the same real-world environments and angle, highlighting YOLO-SMUG alongside other models. The images distinctly show that YOLO-SMUG demonstrates better detection accuracy and a lower false negative rate compared to the initial model.



(**a**)F-RCNN (**b**)YOLOv5n (**c**)YOLOv5s (**d**)YOLOv6n

(**e**)YOLOv6s (**f**)YOLOv8n (**g**)YOLOv8s (**h**)Ours

Figure 13. Visualization of comparative detection results across algorithms for the same scenario and angle.

4. Discussion

The proposed YOLO-SMUG model significantly enhances the accuracy and efficiency of detecting small infrared objects in drone aerial imagery. Ablation studies validate the effectiveness of each enhancement component, including the Shuffle_Block, MSDA attention mechanism, inverted bottleneck C2f_UIB structure, GhostConv module, and MPDIoU loss function.These components collectively contribute to superior detection performance in infrared environments characterized by high noise, low resolution, and weak edge information. Meanwhile, the lightweight nature of the model facilitates its deployment on practical mobile platforms such as drones.

Furthermore, compared with the methods proposed by Ma et al. [19], Yang et al. [20], and Fang et al. [21], which effectively enhance feature extraction, they often require extensive preprocessing or computationally expensive neural models. In contrast, YOLO-SMUG integrates the Shuffle-Block and MSDA attention mechanisms, directly enhancing feature extraction and noise suppression within the detection pipeline. This approach eliminates the need for explicit background removal or additional residual learning tasks, making the model more efficient.

Similarly, compared with the methods of Zhang et al. [22] and Xu et al. [23], which have demonstrated effectiveness, transformer-based models often suffer from high computational complexity, limiting their deployment on resource-constrained UAV platforms. YOLO-SMUG addresses this issue by leveraging the GhostConv module and the inverted bottleneck C2f_UIB structure to enhance feature fusion and expand the receptive field. As a result, it achieves comparable detection accuracy with significantly lower computational cost, making it more suitable for effective UAV applications while maintaining strong performance in detecting small, low-resolution objects.

Additionally, compared with the approaches of Pan et al. [26] and Zhang et al. [27], which effectively improve edge and texture representation, they often require additional training phases or auxiliary networks, increasing computational overhead. In contrast, YOLO-SMUG utilizes the MPDIoU loss function to enhance localization accuracy, particularly for objects with weak texture details. This strategy directly optimizes bounding box prediction without relying on auxiliary restoration models, resulting in improved efficiency and enhanced detection performance in edge-degraded scenarios.

Comprehensive evaluations demonstrate that YOLO-SMUG outperforms state-of-the-art object detection algorithms, including Faster R-CNN, YOLOv5s, YOLOv5n, YOLOv6s, YOLOv6n, YOLOv8s, and YOLOv8n. Additionally, our model offers a lightweight yet powerful solution suitable for efficient UAV infrared detection deployment.

In the future, the proposed improvements enhance model generalization across diverse environments, making YOLO-SMUG highly applicable to critical drone-based tasks such as wildlife monitoring, search and rescue, and military reconnaissance.

5. Conclusions

This study introduces YOLO-SMUG, a lightweight and efficient model designed for infrared object detection in drone-based applications. By integrating the Shuffle_Block module, MSDA attention mechanism, UIB-Ghost architecture, and MPDIoU loss function, the model significantly enhances detection accuracy while reducing computational complexity. Experimental results show a 4.02% improvement in accuracy and a 3.58% increase in mAP_{50} compared to the baseline, along with a 64.38% reduction in parameters and a 57.04% decrease in computational cost, making it highly suitable for UAV-based infrared detection.

Despite these advancements, some limitations remain. The model's generalizability across diverse datasets has yet to be fully assessed, and its robustness under challenging conditions—such as low illumination, extreme weather, or severe occlusion—requires further validation.

Future research will focus on expanding the evaluation scope by testing YOLO-SMUG on a broader range of drone-acquired datasets to improve its adaptability across different environments. Additionally, the feasibility of multi-drone collaborative detection will be explored to enhance performance in scenarios involving occlusions or partially visible objects. These improvements will further expand the model's applicability in surveillance, search and rescue, environmental monitoring, and other critical domains.

Author Contributions: Conceptualization, X.L.; Methodology, X.L.; Software, X.L.; Validation, X.L.; Data curation, X.L.; Writing—original draft, X.L.; Writing—review & editing, X.L.; Visualization, X.L.; Supervision, X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China, grant No. 62461038.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

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

References

1. Thiels, C.A.; Aho, J.M.; Zietlow, S.P.; Jenkins, D.H. Use of unmanned aerial vehicles for medical product transport. *Air Med. J.* **2015**, *34*, 104–108. [CrossRef] [PubMed]
2. Martinez-Alpiste, I.; Golcarenarenji, G.; Wang, Q.; Alcaraz-Calero, J.M. Search and rescue operation using UAVs: A case study. *Expert Syst. Appl.* **2021**, *178*, 114937. [CrossRef]
3. Lu, L.; Chen, Z.; Wang, R.; Liu, L.; Chi, H. Yolo-inspection: Defect detection method for power transmission lines based on enhanced YOLOv5s. *J. Real-Time Image Process.* **2023**, *20*, 104. [CrossRef]
4. Ollero, A.; Merino, L. Unmanned aerial vehicles as tools for forest-fire fighting. *For. Ecol. Manag.* **2006**, *234*, S263. [CrossRef]
5. Velusamy, P.; Rajendran, S.; Mahendran, R.K.; Naseer, S.; Shafiq, M.; Choi, J.G. Unmanned Aerial Vehicles (UAV) in precision agriculture: Applications and challenges. *Energies* **2021**, *15*, 217. [CrossRef]
6. Zhang, Z.; Zhu, L. A Review on Unmanned Aerial Vehicle Remote Sensing: Platforms, Sensors, Data Processing Methods, and Applications. *Drones* **2023**, *7*, 398. [CrossRef]
7. Kotlinski, M.; Calkowska, J.K. U-Space and UTM deployment as an opportunity for more complex UAV operations including UAV medical transport. *J. Intell. Robot. Syst.* **2022**, *106*, 12. [CrossRef]
8. Polukhin, A.; Gordienko, Y.; Jervan, G.; Stirenko, S. Object detection for rescue operations by high-altitude infrared thermal imaging collected by unmanned aerial vehicles. In Proceedings of the Iberian Conference on Pattern Recognition and Image Analysis, Alicante, Spain, 27–30 June 2023; Springer Nature Switzerland: Cham, Switzerland, 2023; pp. 490–504.

9. Liao, K.C.; Lu, J.H. Using UAV to detect solar module fault conditions of a solar power farm with IR and visual image analysis. *Appl. Sci.* **2021**, *11*, 1835. [CrossRef]
10. Ma, Y.; Wei, K.; Liu, F. Research on Visual Algorithm for Fire Detection of Firefighting UAVs Based on Infrared Imaging. In Proceedings of the International Conference on the Efficiency and Performance Engineering Network, Qingdao, China, 8–11 May 2024; Springer Nature: Cham, Switzerland, 2024; pp. 121–131.
11. Messina, G.; Modica, G. Applications of UAV thermal imagery in precision agriculture: State of the art and future research outlook. *Remote Sens.* **2020**, *12*, 1491. [CrossRef]
12. Christnacher, F.; Hengy, S.; Laurenzis, M.; Matwyschuk, A.; Naz, P.; Schertzer, S.; Schmitt, G. Optical and acoustical UAV detection. In Proceedings of the Electro-Optical Remote Sensing X, Edinburgh, UK, 26–29 September 2016; Volume 9988, pp. 83–95.
13. Zhang, N.; Donahue, J.; Girshick, R.; Darrell, T. Part-based R-CNNs for fine-grained category detection. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 834–849.
14. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
15. Solimani, F.; Cardellicchio, A.; Dimauro, G.; Petrozza, A.; Summerer, S.; Cellini, F.; Renò; V. Optimizing tomato plant phenotyping detection: Boosting YOLOv8 architecture to tackle data complexity. *Comput. Electron. Agric.* **2024**, *218*, 108728. [CrossRef]
16. 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, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
17. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef] [PubMed]
18. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
19. Ma, G.; Li, W.; Bao, H.; Roberts, N.J.; Li, Y.; Zhang, W.; Yang, K.; Jiang, G. UAV equipped with infrared imaging for Cervidae monitoring: Improving detection accuracy by eliminating background information interference. *Ecol. Inform.* **2024**, *81*, 102651. [CrossRef]
20. Yang, Z.; Lian, J.; Liu, J. Infrared UAV target detection based on continuous-coupled neural network. *Micromachines* **2023**, *14*, 2113. [CrossRef] [PubMed]
21. Fang, H.; Ding, L.; Wang, L.; Chang, Y.; Yan, L.; Han, J. Infrared small UAV target detection based on depthwise separable residual dense network and multiscale feature fusion. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–20. [CrossRef]
22. Zhang, Q.; Zhou, L.; An, J. Real-time recognition algorithm of small target for UAV infrared detection. *Sensors* **2024**, *24*, 3075. [CrossRef] [PubMed]
23. Xu, K.; Song, C.; Xie, Y.; Pan, L.; Gan, X.; Huang, G. RMT-YOLOv9s: An Infrared Small Target Detection Method Based on UAV Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1–5.
24. Niu, K.; Wang, C.; Xu, J.; Yang, C.; Zhou, X.; Yang, X. An improved YOLOv5s-Seg detection and segmentation model for the accurate identification of forest fires based on UAV infrared image. *Remote Sens.* **2023**, *15*, 4694. [CrossRef]
25. Liu, M.; Wang, X.; Zhou, A.; Fu, X.; Ma, Y.; Piao, C. Uav-yolo: Small object detection on unmanned aerial vehicle perspective. *Sensors* **2020**, *20*, 2238. [CrossRef]
26. 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. [CrossRef]
27. Zhang, Y.; Cai, Z. CE-RetinaNet: A channel enhancement method for infrared wildlife detection in UAV images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–12.
28. Wang, Y.; Lu, Q.; Ren, B. Wind Turbine Crack Inspection Using a Quadrotor with Image Motion Blur Avoided. *IEEE Robot. Autom. Lett.* **2023**, *8*, 1069–1076. [CrossRef]
29. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 116–131.
30. Zhang, M.; Wang, Z.; Song, W.; Zhao, D.; Zhao, H. Efficient Small-Object Detection in Underwater Images Using the Enhanced YOLOv8 Network. *Appl. Sci.* **2024**, *14*, 1095. [CrossRef]
31. Jiao, J.; Tang, Y.M.; Lin, K.Y.; Gao, Y.; Ma, A.J.; Wang, Y.; Zheng, W.S. DilateFormer: Multi-Scale Dilated Transformer for Visual Recognition. *IEEE Trans. Multimed.* **2023**, *25*, 8906–8919. [CrossRef]
32. Li, H.; Li, J.; Wei, H.; Liu, Z.; Zhan, Z.; Ren, Q. Slim-neck by GSConv: A lightweight-design for real-time detector architectures. *arXiv* **2022**, arXiv:2206.02424v3.
33. Zhang, X.Y.; Hu, G.R.; Li, P.H.; Cao, X.Y.; Zhang, H.; Chen, J.; Zhang, L.L. Lightweight Safflower Recognition Method Based on Improved YOLOv8n. *Acta Agric. Eng.* **2024**, *40*, 163–170.
34. Cao, J.; Bao, W.; Shang, H.; Yuan, M.; Cheng, Q. GCL-YOLO: A GhostConv-based lightweight yolo network for UAV small object detection. *Remote Sens.* **2023**, *15*, 4932. [CrossRef]

35. Liu, L.; Huang, K.; Li, Y.; Zhang, C.; Zhang, S.; Hu, Z. Real-time pedestrian recognition model on edge device using infrared vision system. *J. Real-Time Image Process.* **2025**, *22*, 1–11.
36. Qin, D.; Leichner, C.; Delakis, M.; Fornoni, M.; Luo, S.; Yang, F.; Wang, W.; Banbury, C.; Ye, C.; Akin, B.; et al. MobileNetV4—Universal Models for the Mobile Ecosystem. *arXiv* **2024**, arXiv:2404.10518.
37. Zhao, X.; Zhang, W.; Zhang, H.; Zheng, C.; Ma, J.; Zhang, Z. ITD-YOLOv8: An infrared target detection model based on YOLOv8 for unmanned aerial vehicles. *Drones* **2024**, *8*, 161. [CrossRef]
38. Liu, S.; Cao, L.; Li, Y. Lightweight pedestrian detection network for UAV remote sensing images based on strideless pooling. *Remote Sens.* **2024**, *16*, 2331. [CrossRef]
39. Zhang, H.; Li, G.; Wan, D.; Wang, Z.; Dong, J.; Lin, S.; Deng, L.; Liu, H. DS-YOLO: A dense small object detection algorithm based on inverted bottleneck and multi-scale fusion network. *Biomim. Intell. Robot.* **2024**, *4*, 100190. [CrossRef]
40. Song, K.; Wen, H.; Ji, Y.; Xue, X.; Huang, L.; Yan, Y.; Meng, Q. SIA: RGB-T salient object detection network with salient-illumination awareness. *Opt. Lasers Eng.* **2024**, *172*, 107842. [CrossRef]
41. Yang, W.; He, Q.; Li, Z. A lightweight multidimensional feature network for small object detection on UAVs. *Pattern Anal. Appl.* **2025**, *28*, 29. [CrossRef]
42. Liu, L.; Li, P.; Wang, D.; Zhu, S. A wind turbine damage detection algorithm designed based on YOLOv8. *Appl. Soft Comput.* **2024**, *154*, 111364. [CrossRef]
43. Zhang, Y.; Zhang, H.; Huang, Q.; Han, Y.; Zhao, M. DsP-YOLO: An anchor-free network with DsPAN for small object detection of multiscale defects. *Expert Syst. Appl.* **2024**, *241*, 122669. [CrossRef]
44. Aibibu, T.; Lan, J.; Zeng, Y.; Lu, W.; Gu, N. Feature-enhanced attention and dual-gelan net (feadg-net) for uav infrared small object detection in traffic surveillance. *Drones* **2024**, *8*, 304. [CrossRef]
45. 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]
46. Wang, S.; Jiang, H.; Li, Z.; Yang, J.; Ma, X.; Chen, J.; Tang, X. Phsi-rtdetr: A lightweight infrared small target detection algorithm based on UAV aerial photography. *Drones* **2024**, *8*, 240. [CrossRef]
47. Zhao, X.; Xia, Y.; Zhang, W.; Zheng, C.; Zhang, Z. YOLO-ViT-based method for unmanned aerial vehicle infrared vehicle target detection. *Remote Sens.* **2023**, *15*, 3778. [CrossRef]