electronics

MDPI

Article

An Infrared Aircraft Detection Algorithm Based on Context Perception Feature Enhancement

Gang Liu [1,*], Jiangtao Xi [2], Jun Tong [2] and Hongpeng Xu [1]

[1] College of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China; 13083656032@163.com

[2] School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Wollongong, NSW 2522, Australia; jiangtao@uow.edu.au (J.X.); jtong@uow.edu.au (J.T.)

* Correspondence: gliu@haust.edu.cn

Abstract: To address the issue of insufficient extraction of target features and the resulting impact on detection performance in long-range infrared aircraft target detection caused by small imaging area and weak radiation intensity starting from the idea of perceiving target context to enhance the features extracted by convolutional neural network, this paper proposes a detecting algorithm based on AWFGLC (adaptive weighted fusion of global–local context). Based on the mechanism of AWFGLC, the input feature map is randomly reorganized and partitioned along the channel dimension, resulting in two feature maps. One feature map is utilized by self-attention for global context modeling, establishing the correlation between target features and background features to highlight the salient features of the target, thereby enabling the detecting algorithm to better perceive the global features of the target. The other feature map is subjected to window partitioning, with max pooling and average pooling performed within each window to highlight the local features of the target. Subsequently, self-attention is applied to the pooled feature map for local context modeling, establishing the correlation between the target and its surrounding neighborhood, further enhancing the weaker parts of the target features, and enabling the detecting algorithm to better perceive the local features of the target. Based on the characteristics of the target, an adaptive weighted fusion strategy with learnable parameters is employed to aggregate the global context and local context feature maps. This results in a feature map containing more complete target information, enhancing the ability of the detection algorithm to distinguish between target and background. Finally, this paper integrates the mechanism of AWFGLC into YOLOv7 for the detection of infrared aircraft targets. The experiments indicate that the proposed algorithm achieves mAP50 scores of 97.8% and 88.7% on self-made and publicly available infrared aircraft datasets, respectively. Moreover, the mAP50:95 scores reach 65.7% and 61.2%, respectively. These results outperform those of classical target detection algorithms, indicating the effective realization of infrared aircraft target detection.

Keywords: infrared aircraft; target detection; global context; local context; adaptive weighted fusion

check for updates

Citation: Liu, G.; Xi, J.; Tong, J.; Xu, H. An Infrared Aircraft Detection Algorithm Based on Context Perception Feature Enhancement. *Electronics* **2024**, *13*, 2695. https://doi.org/10.3390/electronics13142695

Academic Editor: Valentina E. Balas

Received: 2 June 2024
Revised: 4 July 2024
Accepted: 6 July 2024
Published: 10 July 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

ATR (automatic target recognition) is one of the key technologies in infrared imaging guidance systems. Generally, it is necessary to detect, identify, and track targets such as aircraft at long distances. In such situations, the target has a small area and weak radiation characteristics on the imaging plane, which makes the existing detection algorithms unable to fully extract the features of infrared target such as aircraft, resulting in poor detection performance. Currently, infrared aircraft target detection mainly relies on traditional methods that accurately model target features. However, precise modeling depends on predefined rules and prior knowledge. Once the operational environment exceeds the preset rules, traditional methods become ineffective. Deep learning network simulates the neural connectivity structure of the human brain, possessing powerful capabilities for

autonomous learning and representation of target features. This approach is an effective means of addressing the problem of infrared aircraft target detection. Target detection methods based on deep learning are mainly divided into region-based two-stage detection algorithms, such as the R-CNN series and regression-based single-stage detection algorithms such as the YOLO series, SSD series, etc. Compared to two-stage algorithms, single-stage algorithms offer faster detection speeds, meeting the real-time requirements of detection [1].

The YOLO series algorithms directly regress the position coordinates and categories of targets, enabling high detection accuracy and fast detection speed with this end-to-end detection approach [2–8]. Compared to other algorithms in the YOLO series, YOLOv5 and YOLOv7 strike a good balance between detection efficiency and accuracy. Therefore, many researchers apply YOLOv5 and YOLOv7 to address practical problems. In terms of practical system applications, References [9] and [10] respectively design person detection algorithms for UAV (unmanned aerial vehicle) monitoring systems based on YOLOv5 and YOLOv7. Reference [11] combines active thermography with YOLOv5 to achieve surface defect detection. In terms of improving detection algorithms, scholars from various countries have conducted extensive research in various aspects such as expanding receptive fields [12], dense connection structure [13], attention mechanism [14–16], feature fusion [17–21], lightweight model [22,23], training strategy and loss [17,24,25], activation function [26], and so on. In reference [12], super-resolution reconstructed images are used as input for the detection algorithm. It integrates a multi-level receptive field structure into the feature fusion network of YOLOv5, enabling the detection algorithm to preserve the original information of small infrared targets. In Reference [13], dense connections are used to add the outputs of each layer of YOLOv5 together, better integrating the positional information of small infrared targets in both shallow and deep networks, thus improving the localization ability of YOLOv5. In reference [14], YOLOv5 serves as the baseline model, and standard rotation-equivariant convolution module and rotation residual module are designed. Through the SE (squeeze-and-excitation) attention mechanism, the importance of each channel is adaptively learned and applied to infrared target recognition for UAV. In Reference [15], SE module and inception structures are embedded in the feature extraction network of YOLOv5 to enhance the ability to extract target features in complex backgrounds. In Reference [16], a multi-scale feature extraction network based on ConvNext is constructed on top of YOLOv5. Then, a coordinate attention module is introduced in the ConvNeXt block to focus on targets and suppress the background. Additionally, a segmentation attention module is introduced in the neck to enhance feature fusion ability. Reference [17] constructs a feature pyramid fusion network that considers adaptive attention and feature enhancement and proposes a loss function based on angle regression to accelerate the convergence speed of model training. After fusion with YOLOv7, it achieves the detection of multi-scale targets in UAV images. In Reference [18], a weighted feature fusion pyramid network based on dilated convolution is proposed. Meanwhile, rotation detection is introduced to obtain more accurate detection boxes and ship navigation direction information. These are integrated into YOLOv7 to achieve the detection of infrared ship targets. In Reference [19], a dual-branch fusion detection network based on YOLOv5 is proposed, which can simultaneously input infrared and visible light images for object detection. In reference [20], a four-channel detection model is constructed using YOLOv5, integrating features from both visible light and infrared images to enhance the detection capabilities for person and vehicle by UAV in forest environments. Reference [21] uses YOLOv5 as baseline and utilizes image features from different modal such as infrared and visible light by constructing a layered residual connection. Based on this, a module of multi-level feature fusion is designed to achieve small object detection in multispectral images. In Reference [22], the MobileViT network is utilized as the feature extraction network for YOLOv7 to reduce the complexity of the detection algorithm and support the recognition of infrared vehicle targets by UAV. In Reference [23], starting from the lightweight YOLOv7-tiny network, improvements are made in terms of attention, receptive field and

feature fusion, leading to the design of the YOLOv7-drone algorithm for recognizing small targets by drones. In Reference [24], based on the YOLOv5 framework, a novel training strategy using domain adaptation method is adopted to address the issue of imbalanced dataset distribution. Additionally, a new loss function based on the Wasserstein Distance is proposed to handle small objects by overcoming the sensitivity problem of intersection over union. In Reference [25], a hard sample mining loss function is utilized during the training of the YOLOv7 to guide the network in enhancing its learning of difficult samples, thus improving the classification capability of YOLOv7 for targets. In Reference [26], the activation function LeakyReLU in YOLOv7 is replaced with SiLU, which improves the convergence speed and generalization of the model during training. Additionally, a targeted detection layer for small objects is designed.

The above methods improve detection algorithms from the perspectives of enlarging receptive fields, dense connection structure, attention mechanism, feature fusion, lightweight design, training strategy, loss function, activation function, etc., but do not utilize the global context information of the target. However, incorporating contextual information into target detection tasks helps distinguish targets from the background, thereby enhancing the performance of detection algorithms. Reference [27] introduces a global context module to model the input feature map, enabling the detection algorithm to better distinguish between target and background. Reference [28] proposes a global context information aggregation module to enhance the global perception capability of the detection algorithm, improving its ability to perceive infrared target and background features under nighttime and occlusion conditions. Reference [29] utilizes multi-scale downsampling to obtain corresponding multi-scale features and selectively integrates spatial context information of multi-scale features, enhancing the perception capability of the model for target spatial position. Reference [30] aggregates contextual information using feature maps of different sizes, thereby enhancing the spatial correlation between the background and foreground. Reference [31] designs a position attention mechanism to capture global context information, thereby completing the task of semantic segmentation in scenes. Reference [32] proposes a lightweight global context modeling framework that is structurally similar to SENet. For the problem of infrared small target detection in complex backgrounds, Reference [33] proposes an attention-guided pyramid context network. This network captures and integrates global context information at different scales to achieve better feature representation. Reference [34] proposes a feature fusion module of a multi-scale receptive field to represent the global context information of deep feature maps, thereby expanding the receptive field of detection network for infrared small target.

References [27–34] utilize a global context mechanism to explore the relationship between target and other objects in the scene, highlighting the salient features of the target and enhancing the discriminative ability of the detection algorithm between target and background features. However, most global context mechanisms are based on modeling global information of the image without emphasizing the local information of the target, resulting in weak modeling capabilities for the target and its surrounding neighborhood.

Local context mechanisms establish the correlation between the target and its surrounding neighborhood, enhancing the weaker features of the target and allowing the detection algorithm to focus more on the local information surrounding the target, thereby enhancing the ability to extract local features of the target. Reference [35] introduces convolution-based local perception units into the Transformer structure, constructing local context to enhance the ability to represent local features of the target. On this basis, it proposes a Swin-transformer-based detector for arbitrarily oriented SAR (Synthetic Aperture Radar) ship detection. Reference [36] proposes a classification algorithm for polarization synthetic aperture radar data based on vision transformer (ViT). This algorithm uses a cellular neural network as the feature extractor and employs local window attention to model the local context of the target, thereby highlighting local features. Reference [37] proposes a multi-level vision transformer model that enhances target features by computing the context of non-overlapping local windows through a self-attention mechanism and applies

it to vision detection task. Reference [38] proposes an attention-based context-aware network (ACANet), which enhances infrared small target features by clustering local features. Reference [39] proposes a deep network framework for infrared small target detection that encompasses feature enhancement, interaction, and comparison. This framework enhances local context information from different feature layers using a patch-based attention module. Inspired by the physical thermal diffusion model of infrared small target and human visual mechanism, Reference [40] designs a dense network of multi-level feature extraction and fusion. This network incorporates Gaussian saliency features and local context features into target feature representation to address the issue of sparse feature extraction.

It is essential to integrate both global and local context and prioritize them based on the characteristics of the target in practical applications. Based on the above analysis, this paper proposes a YOLOv7-based target detection algorithm for infrared aircraft using a novel adaptive weighted fusion mechanism that integrates both global and local context. The proposed mechanism can enhance the feature extraction capability of the detection algorithm for infrared aircraft target and improve the discriminative ability between infrared aircraft target and the background, thereby enhancing the performance of the detection algorithm. The main contributions of this paper are as follows:

(1) The paper introduces a mechanism of adaptive weighted fusion of global–local context (AWFGLC) according to the characteristics of infrared aircraft target, providing more comprehensive target feature information for the detection algorithm.
(2) The AWFGLC-YOLO (adaptive weighted fusion of global–local context—you only look once) algorithm is proposed by integrating the AWFGLC mechanism into different levels of the feature extraction network of YOLOv7. This approach effectively utilizes the physical information from shallow feature maps and the semantic information from deep feature maps. The algorithm is then applied to the detection of infrared aircraft target.

It should be further noted that the mechanism of AWFGLC proposed in this paper is universal and can be applied to other single-stage target detection algorithms. The remaining content of this paper is organized as follows: Section 2 elucidates the core idea and working principle of AWFGLC. Section 3 presents the AWFGLC-YOLO detection framework for infrared aircraft target, which integrates YOLOv7 with AWFGLC. Subsequently, Section 4 conducts experiments on two datasets and analyzes the results(Section 4 conducts experiments on two datasets and analyzes the results). Finally, Section 5 concludes the paper and provides prospects for future research.

2. Adaptive Weighted Fusion of Global-Local Context

The mechanism of AWFGLC can highlight target features, enhance the perception ability for target features, and thereby improve the discriminative ability of the detection algorithm between target and background. The proportion of global and local context in the fused feature map varies based on the characteristics of the target. When the target features are strong, the fused feature map should emphasize global context information, whereas it should focus more on local context when the target features are weaker. The proposed mechanism involves randomly reorganizing and partitioning the input feature map along the channel dimension to obtain two feature maps. One feature map is used for global context modeling. The other feature map is divided into windows and pooled. Then, local context modeling is performed within each pooled window, and the local context feature map is obtained through spatial rearrangement. The adaptive weighted strategy using learnable parameters is employed to aggregate global and local context information, enriching the feature representation and improving the feature extraction capability of the feature extraction network for infrared aircraft target. The overall process of the AWFGLC mechanism is illustrated in Figure 1, which includes five functional units: channel reorganization and division, window partitioning, global context modeling, local context modeling and weighted fusion, as depicted in Figure 2a–e respectively.

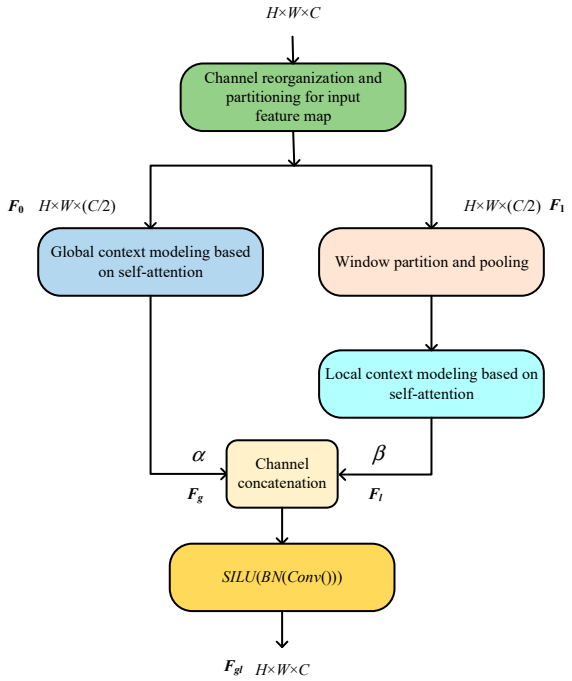


Figure 1. Mechanism of AWFG.

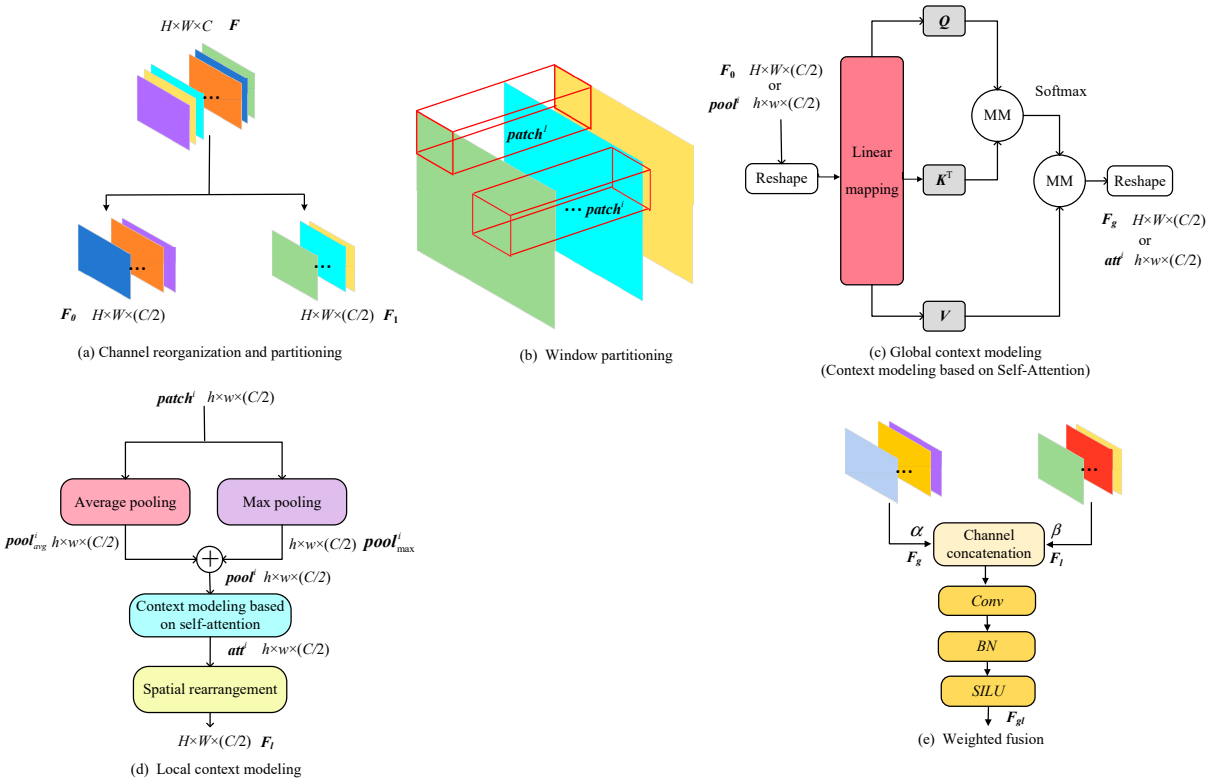


Figure 2. Function units of AWFGLC.

The detailed process of the AWFGLC mechanism is as follows:

(1) Assuming the input feature map is denoted as $\boldsymbol{F} \in R^{H\times W\times C}$. Firstly, $\boldsymbol{F}$ is randomly reorganized along the channel dimension to obtain an intermediate feature map $\boldsymbol{F}' \in R^{H\times W\times C}$. Then, $\boldsymbol{F}'$ is partitioned along the channel dimension to generate feature maps $\boldsymbol{F}_0 \in R^{H\times W\times C/2}$ and $\boldsymbol{F}_1 \in R^{H\times W\times C/2}$, as illustrated in Figure 2a. Random reorganization takes the position index of each channel in feature map $\boldsymbol{F}$, shuffles these indices, and

concatenates the channels according to the shuffled indices. Different random numbers are set to randomly shuffle the input feature map along the channel indices in each epoch during iterative training. This diversifies the arrangement of input feature map. Compared to modeling the context on input feature maps based on specific permutation, modeling the global and local context on feature maps with different permutations helps the detection algorithm learn more diverse and comprehensive features.

To reduce the computational complexity of modeling the context for the input feature map, this paper divides the input feature map equally along the channel dimension and performs separately global and local context modeling. The calculation formulas for random shuffling and partitioning are described as:

$$\boldsymbol{F}'(\boldsymbol{x}_3, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_c, \ldots, \boldsymbol{x}_i) = f_{\mathrm{r}}^{c}(\boldsymbol{F}(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_c)) \tag{1}$$

$$\boldsymbol{F}_0, \boldsymbol{F}_1 = f_{\mathrm{d}}^{2}(\boldsymbol{F}'(\boldsymbol{x}_3, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_c, \ldots, \boldsymbol{x}_i)) \tag{2}$$

where c is the number of channels, f_{r}^{c} represents the channel reorganization of the input feature map, and $\boldsymbol{F}'$ is obtained by changing the orders of $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$, … $\boldsymbol{x}_i$, …$\boldsymbol{x}_c$. f_{d}^{2} is the partitioning for the intermediate feature map, and $\boldsymbol{x_i}$ represents the ith channel of the input feature map. Compared to modeling global and local contexts based on raw feature maps, modeling global and local contexts based on channel partitioning reduces the computational complexity by half.

(2) Global context modeling utilizes self-attention to establish the dependency relationship between pixels in feature map $\boldsymbol{F}_0$ in order to obtain the global context feature map $\boldsymbol{F}_g$, as illustrated in Figure 2c. MM denotes the dot product of vectors. More specifically, the dimension of $\boldsymbol{F}_0$ is transformed into $(C/2) \times (H \times W)$ and mapped to query vector $\boldsymbol{F}_0^q$, key vector $\boldsymbol{F}_0^k$, and value vector $\boldsymbol{F}_0^v$ through linear transformations. The similarity between the query vector $\boldsymbol{F}_0^q$ and the key vector $\boldsymbol{F}_0^k$ is computed using dot product and then divided by $\sqrt{d}$, which is the square root of the input vector's dimension. This ensures that the range of the dot product remains unchanged and reduces the impact of gradient changes, allowing for better backpropagation and optimization. Next, the obtained similarity scores are normalized using the *Softmax* function to obtain attention weights. These weights are then multiplied with the corresponding value vector $\boldsymbol{F}_0^v$, and the resulting vector is reshaped to match the shape of vector $\boldsymbol{F}_0$, yielding the global context feature map $\boldsymbol{F}_g$. The entire calculation process is shown in Equations (3) and (4).

$$\boldsymbol{F}_0^q = f_{Linear}^{q}(\boldsymbol{F}_0), \boldsymbol{F}_0^k = f_{Linear}^{k}(\boldsymbol{F}_0), \boldsymbol{F}_0^v = f_{Linear}^{v}(\boldsymbol{F}_0) \tag{3}$$

$$\boldsymbol{F}_g = Softmax(\frac{\boldsymbol{F}_0^q \cdot \boldsymbol{F}_0^{k^T}}{\sqrt{d}}) \cdot \boldsymbol{F}_0^v \tag{4}$$

In the equations, f_{Linear} represents the linear transformation.

(3) To enhance the local features of the target, a window of $h \times w \times (C/2)$ is slid over $\boldsymbol{F}_1$ to divide it into i patches with dimension of $h \times w \times (C/2)$, as illustrated in Figure 2b. In each patch, max pooling and average pooling are performed to highlight the prominent and average features of the target region, respectively, resulting in the feature map $\boldsymbol{pool}_{\max}^{i}$ and $\boldsymbol{pool}_{\mathrm{avg}}^{i}$. The pooled feature map is obtained by performing matrix addition between $\boldsymbol{pool}_{\max}^{i}$ and $\boldsymbol{pool}_{\mathrm{avg}}^{i}$. The calculation process is illustrated by Equations (5)–(7).

$$\boldsymbol{pool}_{\max}^{i} = f_{\max}^{k \times k \times s}(\boldsymbol{patch}^{i}) \tag{5}$$

$$\boldsymbol{pool}_{avg}^{i} = f_{\mathrm{avg}}^{k \times k \times s}(\boldsymbol{patch}^{i}) \tag{6}$$

$$\boldsymbol{pool}^{i} = \boldsymbol{pool}_{\max}^{i} \oplus \boldsymbol{pool}_{avg}^{i} \tag{7}$$

In the equation, $\boldsymbol{patch}^{i}$ is the ith partitioned window, $f_{avg}^{k \times k \times s}$ represents average pooling performed using convolution with size $k \times k$ and stride s, $f_{\max}^{k \times k \times s}$ denotes max pooling

performed using convolution with size $k \times k$ and stride s, $\oplus$ is matrix addition, and $\boldsymbol{pool}^i$ represents the ith pooled local feature map.

The ith pooled local feature map $\boldsymbol{pool}^i$ is performed with self-attention modeling to adjust the correlation between pixels and obtain $\boldsymbol{att}^i$ which is the pooled local context feature map. This can shift the focus of feature learning more towards the target and its vicinity. Reorganizing $\boldsymbol{att}^i$ in spatial dimension yields the local context feature map $\boldsymbol{F}_l \in R^{H \times W \times (C/2)}$, where the approach of self-attention modeling for each window is similar to $\boldsymbol{F}_0$, as shown in Equations (8)–(10).

$$\boldsymbol{pool}_i^q = f_{Linear}^q(\boldsymbol{pool}^i), \boldsymbol{pool}_i^k = f_{Linear}^k(\boldsymbol{pool}^i), \boldsymbol{pool}_i^v = f_{Linear}^v(\boldsymbol{pool}^i) \tag{8}$$

$$\boldsymbol{att}^i = Softmax(\frac{\boldsymbol{pool}_i^q \cdot \boldsymbol{pool}_i^{k^T}}{\sqrt{d}}) \cdot \boldsymbol{pool}_i^v \tag{9}$$

$$\boldsymbol{F}_l = f_r^s(\boldsymbol{att}^1, \boldsymbol{att}^2, \ldots, \boldsymbol{att}^i) \tag{10}$$

In the equation, f_r^s represents the rearrangement of $\boldsymbol{att}^i$ in the spatial dimension.

(4) With a learnable parameter-based adaptive weighting strategy, the global context feature map $\boldsymbol{F}_g$ and the local context feature map $\boldsymbol{F}_l$ are adaptively weighted and then channel-concatenated. Subsequently, convolution, batch normalization, and activation function are applied to obtain the feature map $\boldsymbol{F}_{gl}$ containing both global and local contextual information. The process is illustrated in Figure 2e. Through the adaptive weighted fusion strategy, the detection algorithm can dynamically adjust the weights of global and local information in $\boldsymbol{F}_{gl}$ based on the characteristics of the infrared aircraft target. When the target features are more prominent, greater emphasis should be placed on the global context, whereas when the target features are less distinct, more focus should be placed on the local context. This enables the detection algorithm to better perceive the surroundings in which the target is located and highlight the target features, thereby enhancing the ability to distinguish between the target and the background. Assuming that the learnable weighting parameters are denoted as α and β, the optimization process involves minimizing the loss function of the target detection algorithm, guiding the optimizer to update these parameters iteratively. The procedure of fusion is represented by Equation (11).

$$\boldsymbol{F}_{gl} = SILU(BN(Conv(f_c(\alpha \cdot \boldsymbol{F}_g, \beta \cdot \boldsymbol{F}_l)))) \tag{11}$$

In the equation, f_c represents channel concatenation, *Conv* denotes a convolutional layer with kernel size and stride both equal to 1×1, *BN* stands for batch normalization, and *SILU* represents the activation function.

The calculation process of Formula (11) will be expanded below. Assuming the convolutional kernel in the convolutional layer is $\boldsymbol{W} = [w_1, w_2, \ldots, w_i, \ldots, w_C]$, the aggregation feature map $\boldsymbol{F}_{agg}$ is obtained by integrating the weighted global and local context feature map, as shown in Equation (12). The aggregation method of contextual features for each channel feature map is shown in Equation (13). The *BN* layer normalizes the aggregation feature map to make its data distribution more stable, thereby helping the detection algorithm to converge faster, as shown in Equation (14). Meanwhile, the detection algorithm can capture and learn the nonlinear expression in the data by applying the activation function, thereby learning higher-level semantic information, as shown in Equation (15).

$$\boldsymbol{F}_{agg}(\boldsymbol{x}_{gl1}, \boldsymbol{x}_{gl2}, \ldots, \boldsymbol{x}_{gli}, \ldots \boldsymbol{x}_{glC}) == \boldsymbol{W} \cdot [\alpha \cdot \boldsymbol{x}_{g1}, \ldots \alpha \cdot \boldsymbol{x}_{gC/2}, \beta \cdot \boldsymbol{x}_{l1}, \ldots, \beta \cdot \boldsymbol{x}_{lC/2}] \tag{12}$$

$$\begin{aligned} \boldsymbol{x}_{gli} = {} & \mathrm{w}_i \cdot \alpha \cdot \boldsymbol{x}_{g1} + \mathrm{w}_i \cdot \alpha \cdot \boldsymbol{x}_{g2} + \ldots + \mathrm{w}_i \cdot \alpha \cdot \boldsymbol{x}_{gC/2} + \\ & w_i \cdot \beta \cdot \boldsymbol{x}_{l1} + w_i \cdot \beta \cdot \boldsymbol{x}_{l2} + \ldots + w_i \cdot \beta \cdot \boldsymbol{x}_{lC/2} \end{aligned} \tag{13}$$

$$\boldsymbol{F}_{agg}^{bn} = \gamma \cdot \frac{\boldsymbol{F}_{agg} - \mu}{\sqrt{\sigma^2 + \varepsilon}} + \delta \tag{14}$$

$$F_{gl} = \frac{F_{agg}^{bn}}{1 + e^{-F_{agg}^{bn}}} \tag{15}$$

In the equation, μ represents the mean of the aggregated feature map F_{agg}, σ is the variance of F_{agg}, γ and δ are the learnable parameters in the *BN* layer, and ε denotes the regularization parameter.

3. The Infrared Aircraft Detection Algorithm Based on Adaptive Weighted Fusion of Global–Local Context

The paper adopts YOLOv7 as the underlying framework and applies the proposed mechanism of adaptive weighted fusion of global–local context to design an infrared aircraft detection algorithm. This algorithm is named AWFGLC-YOLO (adaptive weighted fusion of global–local context—you only look once), and its overall structure is illustrated in Figure 3. The target detection algorithm of YOLOv7 mainly consists of three parts: feature extraction network, feature fusion network and detection head. The feature extraction network consists of basic convolutional module (CBS), extended efficient layer aggregation network (ELAN), and max pooling convolutional module (MP-C3). The feature fusion adopts a network structure of pathway aggregation, which integrates feature information of different scales through top-down and bottom-up approaches. The detection head utilizes the reparameterized convolution (RepConv) to predict results on feature maps, which are downsampled by 8 times, 16 times, and 32 times, respectively.

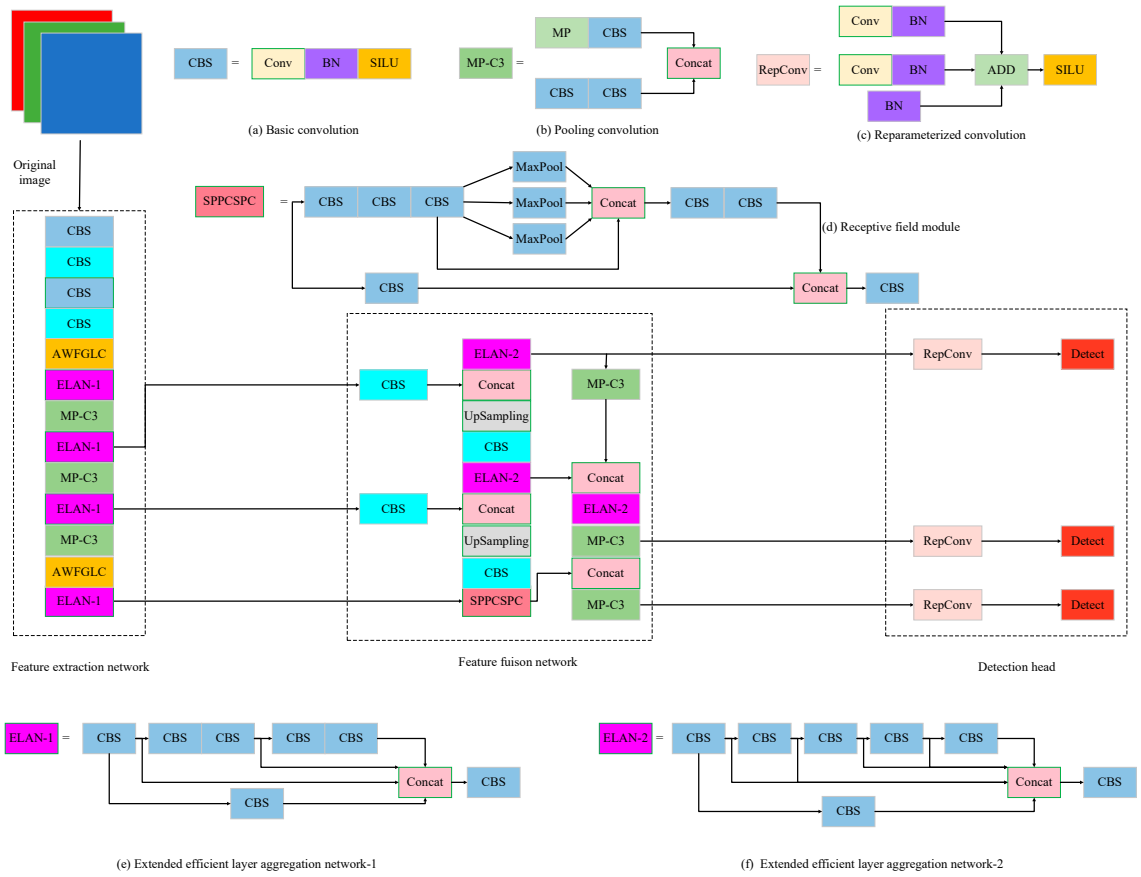


Figure 3. Network architecture of AWFGLC-YOLO.

This paper integrates AWFGLC into the feature extraction network of YOLOv7. It models context information for feature maps which are downsampled by 4 times and 32 times, respectively, fully utilizing the physical information in shallow feature map and the semantic information in deep feature map. This enhances the ability of the detection

algorithm to extract features of infrared aircraft target, thereby improving the discriminative ability between the target and the background.

To verify that the AWFGLC mechanism can highlight target features and improve the ability of the detection algorithm to distinguish between the target and the background, an example for the 4 times downsampled feature map of the feature extraction network is depicted in Figure 4. Figure 4a represents the original image, while Figure 4b represents the input feature map fed into the AWFGLC mechanism. Figure 4c illustrates the global context feature map, and Figure 4d,e respectively depict the window partitioning and pooled feature map as well as the local context feature map. Finally, Figure 4f illustrates the feature map after the AWFGLC mechanism. The content in the red box is an enlarged display of the target area in the image.

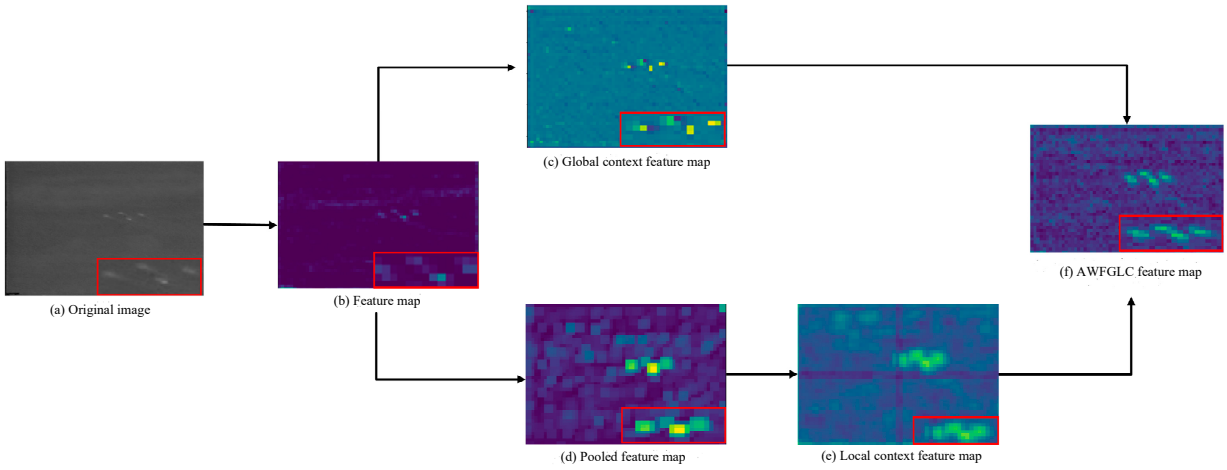


Figure 4. Effect of target feature enhancement.

From Figure 4b, it can be observed that after the feature extraction and downsampling by the first four layers of the feature extraction network, the target features in the input feature map to the AWFGLC mechanism appear weak and relatively small in size. AWFGLC performs channel reorganization and division on the input feature map to obtain two separate feature maps, and then models these two feature maps with different contexts. From Figure 4c, it can be observed that the global context feature map which reflects the correlation between the target and other background pixels in the global image can highlight the brighter tail flame (yellow part) of the target, which helps to distinguish the target from the background. However, the global context feature map has a weaker ability to model the correlation between the target and the neighboring pixels, failing to highlight the darker fuselage of the aircraft. From Figure 4d,e, it can be observed that the feature map of local context based on pooling which reflects the correlation between the target features and the neighboring features can enhance the features of the aircraft fuselage, which helps the detection algorithm better perceive the local features of the target. From Figure 4f, it can be seen that the adaptive weighted fusion strategy effectively aggregates the global context feature map containing brighter tail flame information and the local context feature map with enhanced features of aircraft fuselage, thereby obtaining a feature map containing the overall contour information of the aircraft.

4. Experiment and Analysis

4.1. Experimental Environment

The hardware environment used in this paper includes a CPU with the Intel i7-8700K CPU @ 3.70 GHz, and a total of 32 GB of RAM. The GPU is NVIDIA 3090 Ti, with a memory size of 24 GB. The software environment includes Windows 10 operating system, PyTorch 1.13 deep learning framework, Python 3.7 programming language, and CUDA 11.1 for GPU acceleration.

4.2. Experiment on Self-Made Infrared Dataset

The self-made infrared aircraft dataset consists of 5831 frames, categorized into three poses: back, lateral, and backward. During the detection process, the fuselage and the tail-flame are distinguished simultaneously. Therefore, the manually annotated target categories consist of six classes: back fuselage (BAF), back tail flame (BAT), lateral fuselage (LAF), lateral tail flame (LAT), backward fuselage (BWF), and backward tail flame (BWT). The infrared aircraft dataset is randomly split, with 10% of the data reserved for the test set. The remaining data is subjected to cross-validation, with an 8:2 ratio used to partition the data into training and validation sets.

4.2.1. Experimental Parameters

During the process of model training, the Adam optimizer is employed as the stochastic gradient descent algorithm, with an initial learning rate set to 0.001. The momentum factor is set to 0.937, and the weight decay is set to 0.0005. The initial values for the learnable parameters α and β are both set to 0.5.

The curves showing the changes with the learnable α and β during the training process are depicted in Figure 5. It can be observed from the plot that the values of α and β gradually decrease with the increase of training iterations. According to the criterion of selecting the weights corresponding to the maximum value of the detection metric mAP50:95, the values of α and β are 0.168 and 0.151 respectively. Comparatively, the weight α, which acts on the global context feature map, tends to have slightly larger values than the weight β, which operates on the local context feature map. This indicates that the AWFGLC mechanism is more inclined to learn the global features of the target, especially on the self-made infrared aircraft dataset.

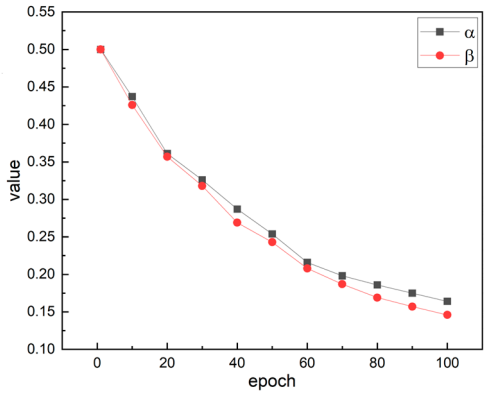


Figure 5. Curves of learnable weight on self-made dataset.

4.2.2. Comparison of Different Contextual Mechanisms

Using YOLOv7 as the benchmark detection algorithm, different contextual mechanisms are added to its feature extraction network at the positions of the 4-times-downsampled feature map and the 32-times-downsampled feature map, including global context (GC) based on simple NLNet [32], global context based on position attention (PA) [31], local context based on local transformer (LT) [37], and the proposed AWFGLC. The experimental results which compare the performance of these different contextual mechanisms in infrared aircraft object detection are shown in Table 1. From Table 1, it can be observed that compared to YOLOv7, the mAP50 of YOLOv7+GC and YOLOv7+LT decreases by 2.1% and 0.9%, while its mAP50:95 decreases by 1.2% and 0.6%, respectively. The mAP50 of YOLOv7+PA increases by 0.3%, with mAP50:95 increasing by 1.2%. The AWFGLC-YOLO detection algorithm, constructed by incorporating the AWFGLC mechanism into YOLOv7, achieves the highest mAP50 and mAP50:95, reaching 97.8% and 65.7%, respectively. This indicates that the AWFGLC mechanism effectively explores both global and local context information in the

input feature map, highlighting target features and enhancing the ability of YOLOv7 to distinguish between the target and the background.

Table 1. Comparative experiment of different context mechanisms.

Method	mAP50/%	mAP50:95/%	Parameter/M	FLOPs/G
YOLOv7	96.4	61.4	35.6	105.4
YOLOv7+GC	94.3	60.2	36.2	105.9
YOLOv7+PA	96.7	62.6	36.1	106.1
YOLOv7+LT	95.5	60.8	36.5	106.7
AWFGLC-YOLO	97.8	65.7	36.9	107.2

To visually observe how the AWFGLC-YOLO algorithm can better extract target features, this study utilizes Eigen-CAM [41] to visualize the features learned by the convolutional layers of different context mechanisms, as shown in Figure 6.

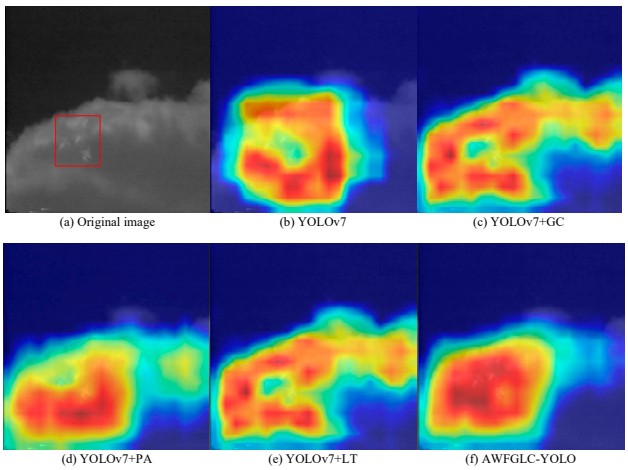


Figure 6. Visualization heatmaps of different context mechanisms.

In Figure 6a, the radiance intensity of the fuselage is similar to that of the surrounding area, indicating limited usable information and partial obstruction of the fuselage by clouds. The content in the red box is the target area in the image. From Figure 6b–e—after incorporating GC, PA, and LT into the feature extraction network of YOLOv7—compared to the original YOLOv7 algorithm, some targets can be observed, but the focus is more on the background area. From Figure 6f, it can be observed that when incorporating the AWFGLC mechanism into the feature extraction network of YOLOv7, the detection algorithm pays more attention to the target and its surrounding neighborhood.

4.2.3. Comparison of Different Detection Algorithms

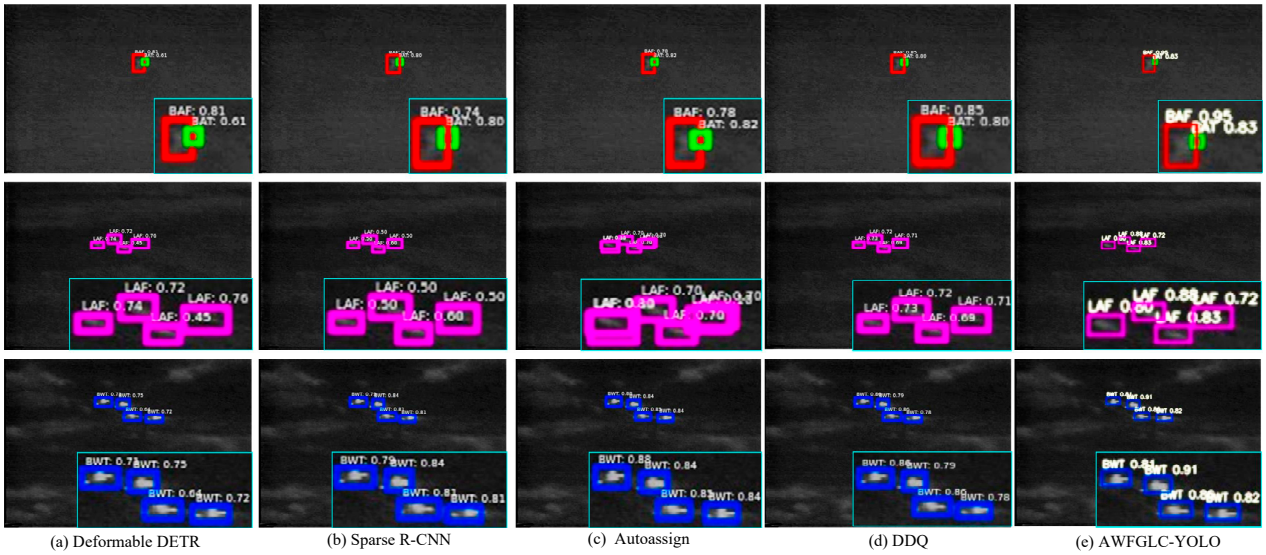
The experimental results comparing our algorithm with CenterNet [42], Sparse R-CNN [43], Efficientdet [44], Autoassign [45], YOLOF [46], Deformable DETR [47], YOLOv5 [3], YOLOv7 [6], YOLOv8 [8], and DDQ [48] are shown in Table 2. It can be seen from this that the mAP50 and mAP50:95 of Efficientdet and YOLOF are relatively low. Sparse R-CNN has good detection performance, with mAP50 and mAP50:95 reaching 90.8% and 59.7%, respectively. However, its model parameter count and floating-point computational complexity are relatively large, reaching 105.9M and 64.6G, respectively. The mAP50 and mAP50:95 of DDQ are 97.2% and 65.3%, respectively, but the computational complexity is relatively large. YOLOv5 demonstrates mAP50 score close to those of CenterNet, Autoassign, and Deformable DETR. However, YOLOv5 achieves mAP50:95 score 1.4% higher than the third-ranked YOLOv8, with smaller model parameters and floating-point computation, at 13.9 M and 16.4 G, respectively. Compared to YOLOv5 and YOLOv8, YOLOv7 achieves a higher mAP50 of 96.4%. However, the mAP50:95 of YOLOv7 is 2.3% lower than that of YOLOv5, and YOLOv7 has larger model

parameters and computational requirements of 35.6 M and 105.4 G, respectively. Compared to other detection algorithms, AWFGLC-YOLO achieves the highest mAP50 and mAP50:95, at 97.8% and 65.7%, respectively. Due to the introduction of the AWFGLC mechanism, the parameter count and floating-point computational requirement of the proposed algorithm increase by 3.7% and 1.7%, respectively, compared to YOLOv7.

Table 2. Comparative experiment of different detection algorithms on self-made dataset.

Method	mAP50/%	mAP50:95/%	Params/M	FLOPs/G
CenterNet	92.4	59.1	20.4	14.2
Autoassign	94.1	60.5	79.1	36.0
Efficientdet	73.7	45.9	18.3	46.4
Sparse R-CNN	90.8	59.7	105.9	64.6
YOLOF	80.6	49.3	42.3	41.2
Deformable DETR	93.8	59.2	40.1	27.4
DDQ	97.2	65.3	47.2	203.9
YOLOv5	94.4	63.7	13.9	16.4
YOLOv7	96.4	61.4	35.6	105.4
YOLOv8	96.0	62.3	25.0	79.1
AWFGLC-YOLO	97.8	65.7	36.9	107.2

The partial visual detection results with different algorithms are shown in Figure 7. The content in the light green box at the bottom right of each image is an enlarged display of the detection result. From the first row of images in Figure 7, it can be seen that Deformable DETR, Sparse R-CNN, Autoassign, DDQ, and AWFGLC-YOLO are all capable of detecting targets. The proposed AWFGLC-YOLO algorithm detects back fuselage (BAF) and back tail flame (BAT) with the highest confidence, reaching values of 0.95 and 0.83, respectively. From the second row of images, it can be observed that Autoassign falsely detects two lateral fuselages (LAF), while other detection algorithms correctly identify the targets. However, AWFGLC-YOLO achieves the highest confidence in detection. As can be seen from the third row of images, Deformable DETR, Sparse R-CNN, Autoassign, and DDQ can correctly detect targets with an average confidence level of 0.710, 0.813, 0.848 and 0.823, respectively. The proposed algorithm in this paper also correctly detects the number, position, and category of the targets with an average confidence reaching the highest value of 0.85.



(a) Deformable DETR (b) Sparse R-CNN (c) Autoassign (d) DDQ (e) AWFGLC-YOLO

Figure 7. Detection results of different algorithms on self-made infrared dataset.

The target size is relatively large, and the contour information is obvious in this dataset. Integrating contextual information, especially global contextual information, into the object detection algorithm improves the detection accuracy of the algorithm. The experimental results confirm the effectiveness of the proposed AWFGLC mechanism.

4.3. Experiment on Infrared Small Target Dataset

The targets in this dataset are UAV on ground or sky backgrounds, with a total of 22 typical scenarios covering single targets on sky backgrounds, two targets in flight crossing on sky backgrounds, single targets on complex ground backgrounds, targets approaching from far to near, targets receding from near to far, and so on [49]. There are 16,177 frames of images and 16,944 targets in the dataset, with most of the target sizes being 3×3 and 5×5 pixels. Small targets typically carry limited feature information and lack distinctive appearance information to differentiate them from the background or similar objects. Incorporating contextual information around the target in natural scene-based target detection tasks can enrich the feature representation, aiding in distinguishing small targets and thereby enhancing the detection performance of the model. The training set, validation set, and test set of infrared small target data adopt the same partitioning strategy as the self-made infrared aircraft dataset. Meanwhile, the experimental parameters of this dataset are the same as those of the self-made infrared aircraft dataset.

The variation curves of weight α and β that can be learned during the training process are shown in Figure 8. From the graph, it can be seen that the values of learnable weights α and β gradually decrease with the increase of training iterations. The optimal weights learned are 0.065 for α and 0.102 for β. Compared to the weight α acting on the global context feature map, the weight β acting on the local context feature map has a larger value. This is because the targets are small in size and have blurred contours in the publicly available dataset, which makes the AWFGLC mechanism more inclined to learn the local features of the targets.

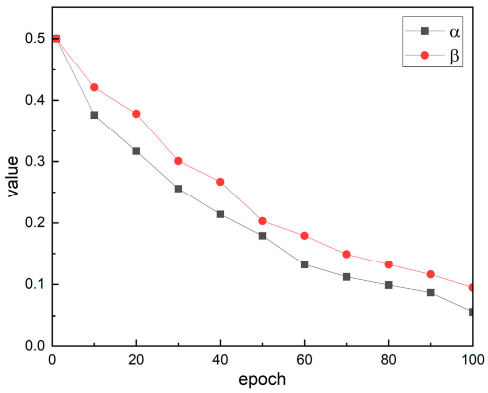


Figure 8. Curves of learnable weights on the public dataset.

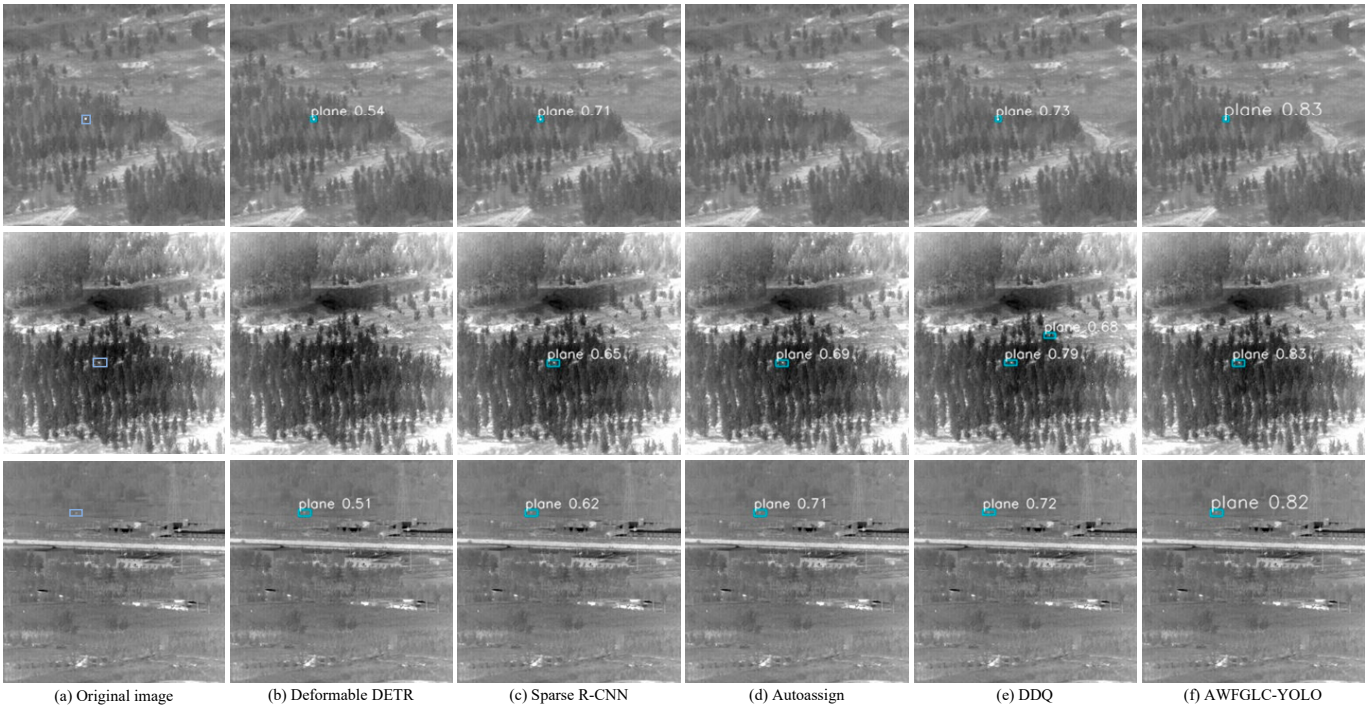
The detection results of different algorithms on the infrared small target dataset on ground or sky backgrounds are shown in Table 3. It can be observed that the mAP50 and mAP50:95 of the proposed algorithm reach 88.7% and 61.2%, respectively, which are higher than those of other detection algorithms.

The partial visual detection results with different algorithms are shown in Figure 9. The content within the blue box of each image represents the region where the target is located.From the first image in Figure 9a, it can be seen that the drone is located above the forest and that its radiation intensity is higher than that of the surrounding background. In the second image of Figure 9a, the drone is located above the forest, and its radiation intensity is similar to that of other background objects in the forest. In the third image of Figure 9a, the drone is located above complex ground backgrounds, such as roads and buildings, with weak target radiation intensity. From the first row of images in Figure 9b–f, it can be seen that except for the Autoassign algorithm, all other detection algorithms can

detect the targets, and the confidence level of AWFGLC-YOLO is the highest, reaching 0.83. In the second row of images, Deformable DETR fails to detect the targets. DDQ detects two targets, with a confidence level of 0.79 for the true target and 0.68 for the false target. Other detection algorithms participating in the comparison are able to detect the targets, and the highest confidence level of AWFGLC-YOLO is 0.83. From the third row of images, it can be seen that all the detection algorithms participating in the comparison can detect the targets. The confidence level of AWFGLC-YOLO is 0.82, which is higher than the confidence level of other detection algorithms.

Table 3. Comparative experiment of different detection algorithms on the public dataset.

Method	mAP50%	mAP50:95%	Params/M	FLOPs/G
CenterNet	79.6	50.8	20.4	14.2
Autoassign	80.0	51.7	79.1	36.0
Efficientdet	64.2	48.1	18.3	46.4
Sparse R-CNN	76.4	50.2	105.9	64.6
YOLOF	54.2	20.4	42.3	41.2
Deformable DETR	81.5	52.2	40.1	27.4
DDQ	87.3	60.4	47.2	203.9
YOLOv5	84.7	57.4	13.9	16.4
YOLOv7	85.3	56.7	35.6	105.4
YOLOv8	85.7	57.2	25.0	79.1
AWFGLC-YOLO	88.7	61.2	36.9	107.2



(a) Original image (b) Deformable DETR (c) Sparse R-CNN (d) Autoassign (e) DDQ (f) AWFGLC-YOLO

Figure 9. Detection results of different algorithms on the public dataset.

In this dataset, the target size is small and the contour information is blurry. Therefore, context information, especially local context information, is integrated into the detection algorithm to improve the ability to distinguish between the target and the background. The experimental results further confirm the role of the context fusion mechanism.

5. Conclusions

In response to the problem of insufficient extraction of target features due to small imaging area and weak radiation intensity in long-distance target detection for infrared

aircraft, which affects detection performance, this paper proposes an infrared aircraft detection algorithm based on adaptive weighted fusion of global–local context (AWFGLC) mechanism. Starting from the idea of perceiving target context to enhance the features extracted by convolutional neural network, this mechanism applies self-attention to model both the global and local contexts of aircraft in infrared image separately. Then, it utilizes an adaptive weighting strategy to fuse the global and local context information based on the characteristics of the target, enhancing the feature representation of infrared aircraft target and improving the ability of the detection algorithm to distinguish between the target and the background. When the target features are more prominent, the AWFGLC mechanism tends to learn the global features of the target, and the fusion result focuses more on the global context. When the target features are weaker, the mechanism tends to learn the local features of the target, and the fusion result focuses more on the local context. Theoretical analysis and experimental results demonstrate that the proposed AWFGLC-YOLO algorithm is effective in detecting infrared aircraft targets. It should be noted that the AWFGLC mechanism proposed in this paper operates on feature maps extracted by convolutional neural networks, making it versatile for application in other single-stage target detection algorithms such as SSD (Single Shot MultiBox Detector).

The deep convolutional neural network has excellent performance, but its high complexity and non-linearity result in low transparency and poor interpretability. Using knowledge to guide the inference process of deep model to make the target detection process interpretable is a future research direction. Therefore, future research efforts could consider modeling knowledge about infrared aircraft radiation, motion, and other factors and integrate them into deep-learning-based target detection algorithms to further improve their performance and interpretability.

6. Patents

The research results of this article have been applied for a national invention patent in China (No. 202410588995.1).

Author Contributions: Conceptualization, G.L.; methodology, G.L., J.X., and J.T.; writing—original draft, G.L.; writing—review and editing, J.X. and J.T.; funding acquisition, J.X.; software, H.X.; validation, H.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the China Scholarship Council (No. [2022]20) and the Key Scientific Research Project of Higher Education Institutions in Henan Province (No. 21A520012).

Data Availability Statement: Data in this article can be downloaded at http://www.dx.doi.org/10.11922/sciencedb.902 (accessed on 17 August 2020).

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

References

1. Cheng, X.; Song, C.; Shi, J.; Zhou, L.; Zhang, Y.; Zheng, Y. A review of general object detection based on deep learning. *Acta Electron. Sin.* **2021**, *49*, 1428–1438.
2. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
3. Jocher, G.; Stoken, A.; Borovec, J.; Changyu, L.; Hogan, A.; Diaconu, L.; Ingham, F.; Poznanski, J.; Fang, J.; Yu, L.; et al. Ultralytics/yolov5: v3.1-Bug Fixes and Performance Improvements. 2020. Available online: https://github.com/ultralytics/yolov5 (accessed on 25 June 2020).
4. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. Scaled-YOLOv4: Scaling cross stage partial network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 13029–13038.
5. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOx: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
6. 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, 18–22 June 2023; pp. 7464–7475.
7. Wang, C.Y.; Yeh, I.H.; Liao, H.Y.M. You only learn one representation: Unified network for multiple tasks. *J. Inf. Sci. Eng.* **2023**, *39*, 691–709.
8. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLO (Version 8.0.0). 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 10 January 2023).

9. Mantau, A.J.; Widayat, I.W.; Leu, J.-S.; Köppen, M. A human-detection method based on YOLOv5 and transfer learning using thermal image data from UAV perspective for surveillance system. *Drones* **2022**, *6*, 290. [CrossRef]
10. Guettala, W.; Sayah, A.; Kahloul, L.; Tibermacine, A. Real time human detection by unmanned aerial vehicles. In Proceedings of the 2022 International Symposium on iNnovative Informatics of Biskra (ISNIB), Biskra, Algeria, 7–8 December 2022; pp. 165–170.
11. Lema, D.G.; Pedrayes, O.D.; Usamentiaga, R.; Venegas, P.; García, D.F. Automated detection of subsurface defects using active ther-mography and deep learning object detectors. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 4503213. [CrossRef]
12. Zhou, X.; Jiang, L.; Hu, C.; Lei, S.; Zhang, T.; Mou, X. YOLO-SASE: An improved YOLO algorithm for the small targets detection in complex backgrounds. *Sensors* **2022**, *22*, 4600. [CrossRef] [PubMed]
13. Wang, K.; Li, S.; Niu, S.; Zhang, K. Detection of infrared small targets using feature fusion convolutional network. *IEEE Access* **2019**, *7*, 146081–146092. [CrossRef]
14. Xiao, F.; Lu, H.; Zhang, W.; Huang, S.; Jiao, Y.; Lu, Z.; Li, Z. Aerial infrared image target recognition algorithm based on rotation equivariant convolution. *Acta Armamentarii* **2023**, 1–9. Available online: https://link.cnki.net/urlid/11.2176.TJ.20231018.1031.004 (accessed on 18 October 2023).
15. Zhou, W.; Wu, Z.; Zhang, Z.; Peng, L.; Xie, L. Lightweight small target detection method based on weak feature enhancement. *Control. Decis.* **2024**, *39*, 381–390.
16. Zhou, J.; Zhang, B.; Yuan, X.; Lian, C.; Ji, L.; Zhang, Q.; Yue, J. YOLO-CIR:The network based on YOLO and ConvNeXt for infrared object detection. *Infrared Phys. Technol.* **2023**, *131*, 104703. [CrossRef]
17. Zhang, H.; Shao, F.; He, X.; Chu, W.; Zhao, D.; Zhang, Z.; Bi, S. ATS-YOLOv7: A real -time multi-scale object detection method for UAV aerial images based on improved YOLOv7. *Electronics* **2023**, *12*, 4886. [CrossRef]
18. Deng, H.; Zhang, Y. FMR-YOLO: Infrared ship rotating target detection based on synthetic fog and multiscale weighted feature fusion. *IEEE Trans. Instrum. Meas.* **2023**, *73*, 5001717. [CrossRef]
19. Hou, Z.; Yang, C.; Sun, Y.; Ma, S.; Yang, X.; Fan, J. An object detection algorithm based on infrared-visible dual modal feature fusion. *Infrared Phys. Technol.* **2024**, *137*, 105107. [CrossRef]
20. Marques, T.; Carreira, S.; Miragaia, R.; Ramos, J.; Pereira, A. Applying deep learning to real-time UAV-based forest monitoring: Leveraging multi-sensor imagery for improved results. *Expert Syst. Appl.* **2024**, *245*, 123107. [CrossRef]
21. Sun, J.; Yin, M.; Wang, Z.; Xie, T.; Bei, S. Multispectral object detection based on multilevel feature fusion and dual feature modulation. *Electronics* **2024**, *13*, 443. [CrossRef]
22. 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]
23. Xue, S.; An, H.; Lv, Q.; Cao, G. Image object detection algorithm based on YOLOv7-tiny in complex background. *Infrared Laser Eng.* **2024**, *53*, 20230472-1–20230472-12.
24. Kim, J.; Huh, J.; Park, I.; Bak, J.; Kim, D.; Lee, S. Small object detection in infrared images: Learning from imbalanced cross-domain data via domain adaptation. *Appl. Sci.* **2022**, *12*, 11201. [CrossRef]
25. Hu, S.; Zhao, F.; Lu, H.; Deng, Y.; Du, J.; Shen, X. Improving YOLOv7-tiny for infrared and visible light image object detection on drones. *Remote. Sens.* **2023**, *15*, 3214. [CrossRef]
26. Zhang, G.; Li, C.; Li, G.; Lu, W. Small target detection algorithm for UAV aerial images based on improved YOLOv7-tiny. *Adv. Eng. Sci.* **2023**, 1–14. [CrossRef]
27. Tan, J.; Yin, W.; Liu, L.; Wang, Y. DenseNet-siamese network with global context feature module for object tracking. *J. Electron. Inf. Technol.* **2021**, *43*, 179–186.
28. Hou, Z.; Sun, Y.; Guo, H.; Li, J.; Ma, S.; Fan, J. M-YOLO: An object detector based on global context information for infrared images. *J. Real-Time Image Process.* **2022**, *19*, 1009–1022. [CrossRef]
29. Wang, Z.; Xu, Z.; Xue, Y.; Lang, C.; Li, Z.; Wei, L. Global and spatial multi-scale contexts fusion for vehicle re-identification. *J. Image Graph.* **2023**, *28*, 471–482.
30. Zhang, W.; Fu, C.; Xie, H.; Zhu, M.; Tie, M.; Chen, J. Global context aware RCNN for object detection. *Neural Comput. Appl.* **2021**, *33*, 11627–11639. [CrossRef]
31. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3141–3149.
32. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. GCNet: Non-local networks meet squeeze-excitation networks and beyond. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1971–1980.
33. 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]
34. Zhang, Y.; Li, C.; Liu, Y.; Liu, Z.; Yang, R. Multi-scale feature fusion attention network for infrared small target detection. In *Fourteenth International Conference on Graphics and Image Processing*; SPIE: Bellingham, WA, USA, 2022; Volume 12705, pp. 34–43.
35. Yang, Z.; Xia, X.; Liu, Y.; Wen, G.; Zhang, W.E.; Guo, L. LPST-Det: Local-perception-enhanced swin transformer for SAR ship detection. *Remote Sens.* **2024**, *16*, 483. [CrossRef]
36. Jamali, A.; Roy, S.K.; Bhattacharya, A.; Ghamisi, P. Local window attention transformer for polarimetric SAR image classification. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 2735–2751. [CrossRef]

37. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
38. Ling, S.; Chen, L.; Wu, Y.; Zhang, Y.; Gao, Z. ACANet: Attention-based context-aware network for infrared small target detection. *J. Supercomput.* **2024**, *4*, 1–29. [CrossRef]
39. Lv, G.; Dong, L.; Xu, W. Hierarchical interactive multi-granularity co-attention embedding to improve the small infrared target detection. *Appl. Intell.* **2023**, *53*, 27998–28020. [CrossRef]
40. Ma, T.; Yang, Z.; Fan, Y.; Song, Y.-F.; Liang, J.; Wang, H. DMEF-Net: Lightweight infrared dim small target detection network for limited samples. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5626015. [CrossRef]
41. Muhammad, M.B.; Yeasin, M. Eigen-cam: Class activation map using principal components. In Proceedings of the 2020 International Joint Conference on Neural Networks, Glasgow, UK, 19–24 July 2020; pp. 2131–2147.
42. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6569–6578.
43. Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.; Wang, C.; et al. Sparse R-CNN: End-to-end object detection with learnable proposals. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 14454–14463.
44. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
45. Zhu, B.; Wang, J.; Jiang, Z.; Zong, F.; Liu, S.; Li, Z.; Sun, J. Autoassign: Differentiable label assignment for dense object detection. *arXiv* **2020**, arXiv:2007.03496.
46. Chen, Q.; Wang, Y.; Yang, T.; Zhang, X.; Cheng, J.; Sun, J. You only look one-level feature. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 13039–13048.
47. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable transformers for end-to-end object detection. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020; pp. 12254–12272.
48. Zhang, S.; Wang, X.; Wang, J.; Pang, J.; Lyu, C.; Zhang, W.; Luo, P.; Chen, K. Dense distinct query for end-to-end object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 7329–7338.
49. Hui, B.; Song, Z.; Fan, H.; Zhong, P.; Hu, W.; Zhang, X.; Ling, J.; Su, H.; Jin, W.; Zhang, Y.; et al. Dataset of weak small aircraft detection and tracking in ground/air background infrared images. *China Sci. Data* **2020**, *5*, 291–302.

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.