World Scientific Research Journal Volume 11 Issue 12, 2025
ISSN: 2472-3703 DOI: 10.6911/WSRJ.202512_11(12).0007

DAP-YOLO: A Multi-scale Adaptive Fusion Human Detection Model for Aerial Photography Disaster Scenarios

Shengmin Zhu [a], Tingting Geng [b,*], Bingqian Ji [c], Qi Liu [d]

School of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo 454000, Henan, China

[a]212304020095@home.hpu.edu.cn, [b]212204020017@home.hpu.edu.cn, [c]212304020092@home.hpu.edu.cn, [d]212304020025@home.hpu.edu.cn

[*] Corresponding author

Abstract

In emergency rescue operations for natural disasters and traffic accidents, rapidly and accurately detecting core disaster areas and identifying survivors in urgent need of rescue holds significant practical importance. To address this, this study proposes a multi-scale adaptive fusion human detection model for aerial disaster scenes—DAP-YOLO. First, the PPA module is introduced, employing multi-branch parallel fusion to enhance focus on small objects. Second, the C3k2_DWR module utilizes DWR to expand the receptive field, mitigating feature information loss during downsampling. The ASFF module is integrated into the detection layer, dynamically assigning weights to adaptively adjust the contribution of feature maps at different scales. Additionally, a micro-detection structure is designed for small objects, significantly reducing model parameters while enhancing detection performance. Experiments demonstrate that DAP-YOLO outperforms YOLO11n on the C2A dataset by 6.70% in recall, 5.30% in mAP50, 5.10% in mAP50-95, and 4.63% in F1-Score. and on the SARD dataset, improvements of 6.40%, 4.10%, 8.20%, and 4.78% respectively. Further comparison of mAP50-95 reveals that DAP-YOLO outperforms YOLOv12n by 5.10% while significantly surpassing the lightweight YOLOv13n model. These experiments validate the effectiveness of human detection models in disaster scenarios, providing technical support for drone-based post-disaster search and rescue operations.

Keywords

Aerial disaster imagery; PPA; Adaptive Spatial Feature Fusion; Tiny human body human target; Dilation wise residual; YOLO11.

1. INTRODUCTION

Despite significant and continuous advances in science and technology, disaster and crisis management continues to pose a vast and complex challenge [1]. In disaster sites, rapid and accurate human detection is vital for locating survivors and supporting rescue operations. Drones, with their high efficiency, play a key role in reducing casualties. However, in diverse disaster conditions—such as floods, fires, and structural collapses—drones must capture images from various angles and altitudes. These viewpoint changes cause targets to appear at different scales and shapes, increasing detection uncertainty. [2]. Moreover, drone images often contain small, unevenly distributed, or blurred targets due to motion, further degrading detection accuracy [3]. Overall, the interplay of multiple complex factors in disaster scenarios severely disrupts the stability and accuracy of human target detection tasks in aerial imagery.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202512_11(12).0007

At present, object detection algorithms are primarily categorized into two-stage and single-stage approaches. The CNN-based R-CNN series exemplifies typical two-stage methods, while the YOLO series [4]and SSD [5]serve as mainstream single-stage detectors. YOLO has emerged as the most widely adopted single-stage framework due to its end-to-end prediction paradigm, computational efficiency, and strong real-time performance. With advances in deep learning, Transformer architecture has catalyzed the evolution of detection models. For instance, DETR [6] introduced an end-to-end prediction paradigm to enhance global modeling capabilities, Deformable DETR [7], Swin Transformer [8], and PVT [9] have improved multi-scale modeling by leveraging deformable attention and hierarchical pyramid mechanisms.

2. RELATED RESEARCH

In recent years, advances in object detection have driven extensive research on human detection across various fields, including video surveillance, robotics, autonomous driving, and aerial imaging. To tackle scenario-specific challenges, existing studies have focused on occlusion handling, small-object detection, multi-scale feature fusion, and lightweight model design. Regarding occlusion and complex background modeling, Jin et al. [10] proposed IF-RCNN to reduce background interference and pedestrian blurring in tunnels, significantly improving detection accuracy. Zhang et al. [11] enhanced feature–semantic correlation by integrating pedestrian attributes, improving performance under occlusion and complex poses. Wang et al. [12] introduced Wise-IoU and BiFormer attention into YOLOv8 to enhance feature fusion in occluded scenarios. In aerial disaster applications, Alotaibi et al. [13] developed the LSA algorithm employing multi-UAV coordination for victim search; however, experiments were limited to simulated environments. Qiu et al. [14] improved detection via radio-assisted sensing, though its outdoor robustness remains constrained. Wang et al. proposed MER-YOLO, which mitigates false positives in dense scenes through regional information redistribution. Zhang et al. [15] further enhanced YOLOv8 for small-object detection under occlusion. Du et al. [16] applied sparse convolutions to simplify detection heads, reducing complexity with minor accuracy loss. MS-YOLOv7 [17] incorporates CBAM, Swin Transformer, and SPPFS modules to strengthen small-object perception by integrating attention mechanisms and multi-scale fusion. DCLANet [18], developed on YOLOv5, employs dense cropping and local attention mechanisms to amplify small-target regions, significantly improving small-human detection performance on datasets such as VisDrone2019. In summary, while existing methods have made progress in urban surveillance, crowd detection, and general aerial detection tasks, dedicated research on casualty detection and rescue in disaster scenarios remains limited. Significant challenges persist, particularly in scenarios involving small targets, occlusion, and complex background environments.

In drone-based disaster imagery, human detection faces three main challenges: (1) Dense small targets and occlusion: Victims are often densely distributed with varied postures (prone, lateral, curled), while debris, fire, or fog frequently cause occlusions, reducing detection accuracy. (2) Complex backgrounds: Disaster scenes contain large amounts of non-target information, such as debris, collapsed structures, smoke, and water reflections, which easily confuse detection models. (3) Small-target feature loss: Multi-layer downsampling weakens or removes key features of small objects, limiting traditional detectors' ability to handle scale variations and increasing missed detections.

To overcome these challenges, this study proposes DAP-YOLO, a multi-scale adaptive fusion model for human detection in disaster scenarios based on YOLO11. The main contributions are as follows: (1) The backbone network incorporates a PPA module, utilizing parallel multi-branch fusion and an efficient attention mechanism to extract fine-grained features and enhance focus on small objects. (2) To reduce small object feature loss from downsampling, the

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202512_11(12).0007

original C3k2 module is replaced with C3k2_DWR, which integrates residual attention and multi-scale receptive fields. This enhances feature representation and improves small-object detection. (3) To handle occlusion and complex backgrounds, an ASFF mechanism is applied across multiple scales. It dynamically weights effective features and suppresses conflicts, improving detection accuracy under occlusion. (4) The large-scale detection head is replaced with an upsampling layer and a micro-detection head. The upsampling layer retains fine details and produces high-resolution features, improving detection accuracy while reducing model complexity.

3. MATERIALS AND METHODS

3.1. YOLO11 Detection Algorithm

The YOLO11 model has achieved notable advancements in computer vision. Building upon the strengths of its predecessors, it introduces new modules and optimized strategies that enhance both detection accuracy and inference speed, making it particularly suitable for real-time human detection in disaster scenarios. As illustrated in Figure 1, the network architecture consists of three main components: a backbone network, a neck region, and a detection head. Compared with the representative YOLOv8 framework, YOLO11 incorporates several architectural refinements. Specifically, the C3k2 module replaces the C2f module in the backbone, while the SPPF module is retained for efficient multi-scale feature aggregation. In addition, the C2PSA module, which integrates a spatial attention mechanism, is introduced to further strengthen feature extraction and fusion.

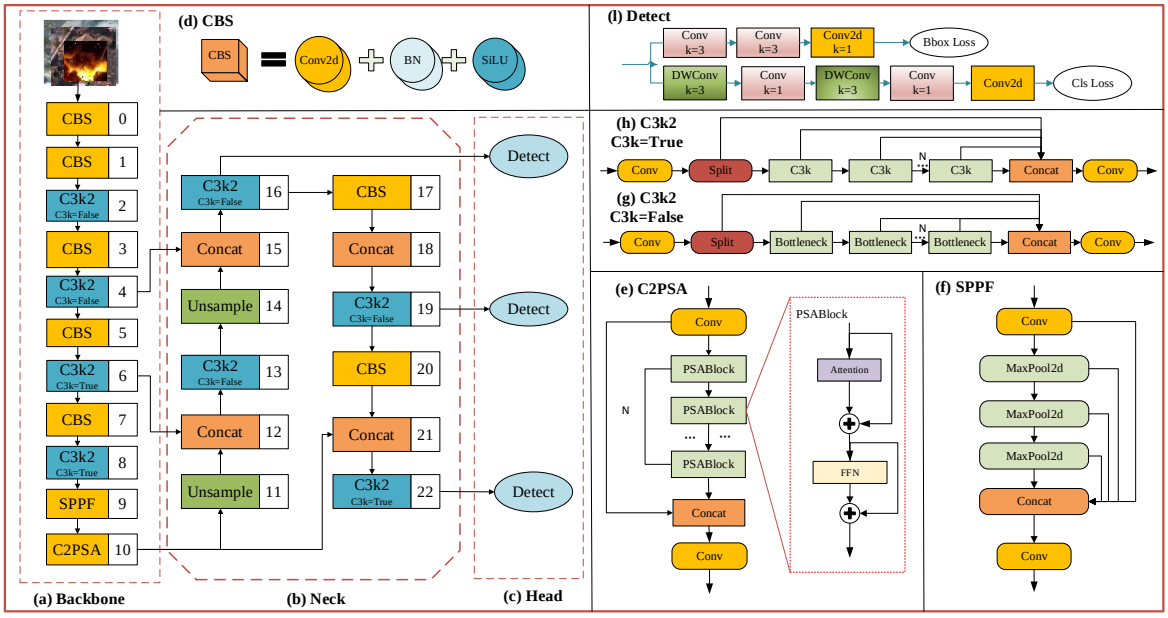


Figure 1. YOLO11 network structure diagram

(1) Backbone Network

The backbone architecture of YOLO11 begins with a sequence of CBS layers, employing a five-stage downsampling strategy. This progressively reduces the spatial resolution of feature maps while increasing their channel depth (Figure 1(a)). The CBS module performs nonlinear transformation and normalization on input features through convolution, batch normalization (BN), and the SiLU activation function (Figure 1(d)). For cross-stage feature integration, the C3k2 module serves as the core feature extraction component, introducing variable convolution kernels and channel separation enhancements compared to traditional C3 modules. The module splits input features into two branches: one branch passes features directly through a convolution layer, while the other processes feature via the C3k module. When C3k=True, this

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202512_11(12).0007

branch employs the C3k module (Figure 1(h)); when C3k = False, N Bottleneck modules extract deep features (Figure 1(g)). The outputs of both branches are subsequently fused through cross-dimensional concatenation. The backbone also retains the enhanced SPPF (Spatial Pyramid Pooling Fast Fusion) module, which connects three max-pooling layers to pool input feature maps into fixed-size outputs. This facilitates cross-scale feature integration and enriches feature expressiveness, thereby improving the modeling of object size, shape, and spatial location (Figure 1(f)). Additionally, the backbone incorporates the C2PSA module, whose key component is the PSABlock. By integrating a self-attention mechanism, this module enhances the spatial expressiveness of feature maps, guiding the model to focus on regions most relevant to the object detection task within the image (Figure 1(e)).

(2) Neck

As shown in Figure 1(b), the neck connects the backbone and output layers, integrating multi-scale features and passing them to the detection branches. To enhance feature aggregation, YOLO11 employs the C3k2 module in the neck, which aligns multi-scale features through upsampling and concatenation, preserving semantic connections and improving feature representation.

(3) Detection Head

The YOLO11 detection head maintains the decoupled YOLO architecture, predicting bounding boxes, class probabilities, and confidence scores. As shown in Figure 1(c), it has three branches—small, medium, and large [19]—to handle objects of different sizes. The classification branch uses binary cross-entropy loss, while the bounding box branch combines distributed focal loss and complete IoU. Structurally, two DWConv layers are added to the classification branch to enhance cross-channel feature representation (Figure 1(l)).

3.2. Algorithm Improvements

(1) DAP-YOLO

The YOLO11 series includes five variants—YOLO11n, s, m, l, and x—differing in width and depth to balance performance and speed. YOLO11n is lightweight, fast, and suitable for edge devices, while YOLO11x is more complex for higher accuracy. Considering disaster relief constraints, this study uses YOLO11n as the base model and further optimizes its architecture.

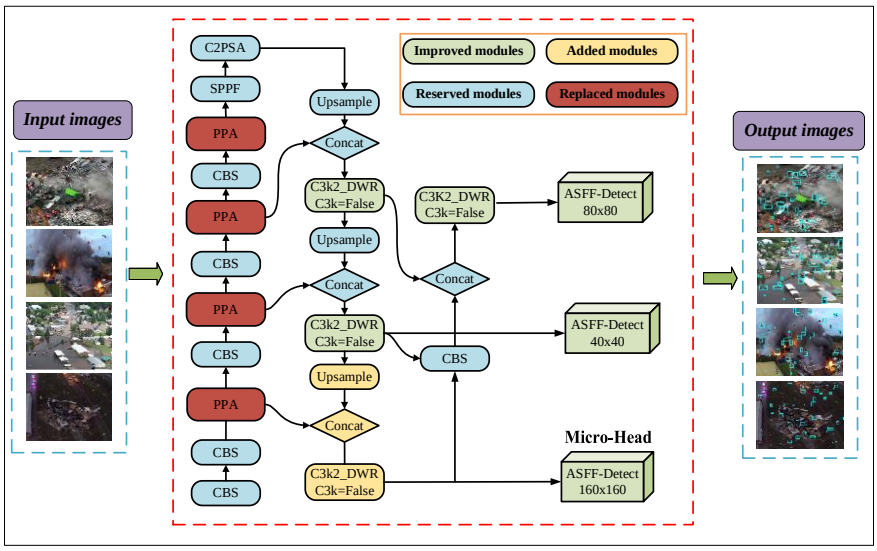


Figure 2. DPA-YOLO network structure diagram

As shown in Figure 2, the model improves three key aspects: feature extraction, scale fusion, and detection architecture. The backbone uses a PPA module to extract detailed features and enhance focus on small objects. The C3k2_DWR module integrates multi-expansion

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202512_11(12).0007

convolutions with residual attention to improve context modeling and perception of occlusions and varying receptive fields. The ASFF-Detect module applies position-aware dynamic weighting for spatially adaptive multi-scale feature fusion, addressing YOLO-based detectors' static limitations. The Micro-Detect architecture removes large-scale detection heads and strengthens the high-resolution branch, reducing complexity while maintaining small-object detection performance and improving edge deployment efficiency.

(2) Parallel Patch-aware Attention Module

To improve small-target detection in disaster scenarios, a Parallel Patch-aware Attention (PPA) module is added to the backbone. As shown in Figure 3, PPA has three branches: local, global, and serial convolution. Input features are first channel-adjusted, then processed in each branch to capture local structures, global dependencies, and multi-layer convolutional features. Branch outputs are fused with weighted attention, and channel and spatial attention mechanisms further enhance task-relevant features. This design improves small-object detection accuracy and robustness while keeping the model lightweight.

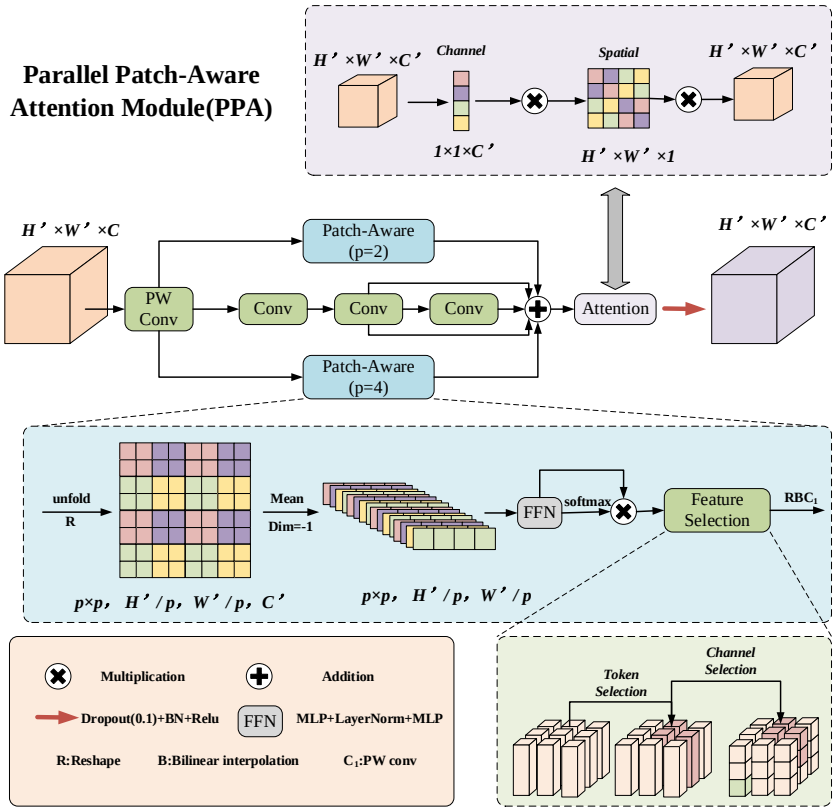


Figure 3. Parallel Patch-aware Attention Module

(3) C3k2_DWR Module

To improve multi-scale context modeling, the Dilation-wise Residual (DWR) mechanism is combined with C3k to form the C3k2_DWR module (Figure 4).

The C3k2_DWR module works as follows. When C3k is False (Figure 4(a)), the input feature map passes through a convolution layer and a Bottleneck structure, enhancing feature learning. The Bottleneck uses convolution and residual connections (Figure 4(b)) and incorporates an expandable residual attention mechanism in the DWR_Conv module. Two branches then fuse local and global information, improving feature representation. DWR_Conv generates refined features via residual attention, batch normalization, and GELU (Figure 4(c)). The DWR core (Figure 4(d)) uses a multi-branch residual architecture to integrate contextual features across different receptive fields. The process is as follows: (1) Input feature maps pass through 3×3 convolutions, BN, and ReLU to extract regional residual features and compress channels. (2) The

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202512_11(12).0007

DWR module applies dilated convolutions with rates d = 1, 3, 5 to expand receptive fields, capture multi-scale context, and enhance feature representation [20]. (3) Outputs from the three dilated convolutions are concatenated to form multi-scale fused features. (4) Residual connections combine input and fused features, mitigating gradient vanishing and improving feature propagation.

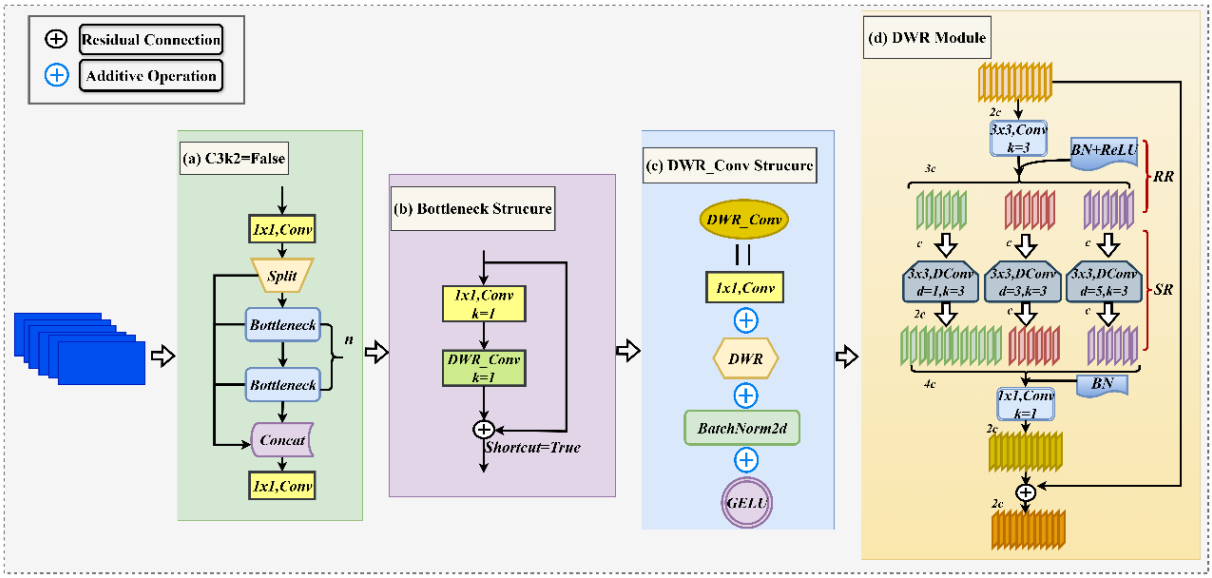


Figure 4. C3k2_DWR module structural decomposition diagram

(4) ASFF-Detect Module

Drone images in disaster scenarios vary greatly in resolution, background, target scale, and density, affecting detection speed and accuracy. To address this, an Adaptive Spatial Feature Fusion (ASFF) mechanism is added to the detection layer (Figure 5(a)) to efficiently fuse multi-scale features. ASFF [21] improves detection by assigning higher attention to small targets and adaptively fusing features across scales, reducing interference from large-scale features.

The ASFF-Detect module performs multi-scale fusion using independent ASFF sub-layers for its three detection heads (Level-1, Level-2, Level-3). As shown in Figure 5(b), features are scaled to match spatial resolution and channel count before fusion. Levels 1 and 2 are upsampled to Level-3 resolution using convolution and bilinear interpolation, while Level-3 is downsampled to Level-1 via stride-2 convolution and max-pooling. The transformed feature maps ($X^{1\to3}$, $X^{2\to3}$, and $X^{3\to3}$) are adaptively fused to produce the output ASFF-3 is expressed as shown in Equation (1):

$$ASFF-3=\alpha^3\cdot X^{1\to3}+\beta^3\cdot X^{2\to3}+\gamma^3\cdot X^{3\to3} \tag{1}$$

Among these, α^3, β^3 and γ^3 are learnable spatial attention weight parameters that control the contribution ratio of feature maps at different scales during fusion, satisfying the normalization constraint $\alpha^3+\beta^3+\gamma^3=1$.

Specifically, y_{ij}^{l} represents the feature vector at position (i,j) within the channel of the output feature map y^{l}, and $x_{ij}^{n\to l}$ represents the feature value at position (i,j) in the feature map transformed from n layer to l layer . This relationship can be further generalized as shown in Equation (2):

$$y_{ij}^{l}=\alpha_{ij}^{l}\cdot x_{ij}^{n\to l}+\beta_{ij}^{l}\cdot x_{ij}^{n\to l}+\gamma_{ij}^{l}\cdot x_{ij}^{n\to l} \tag{2}$$

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202512_11(12).0007

$\lambda^l_{\alpha_{ij}}$, $\lambda^l_{\beta_{ij}}$ and $\lambda^l_{\gamma_{ij}}$ denote the raw attention weights obtained via convolution at position (i, j); α^l_{ij}, β^l_{ij} and γ^l_{ij} represent the spatial attention weights of the feature maps, adaptively fused to the target layer at position x, as implemented in Equations (3)–(5):

$$\alpha^l_{ij} = \frac{e^{\lambda^l_{\alpha_{ij}}}}{e^{\lambda^l_{\alpha_{ij}}} + e^{\lambda^l_{\beta_{ij}}} + e^{\lambda^l_{\gamma_{ij}}}} \tag{3}$$

$$\beta^l_{ij} = \frac{e^{\lambda^l_{\beta_{ij}}}}{e^{\lambda^l_{\alpha_{ij}}} + e^{\lambda^l_{\beta_{ij}}} + e^{\lambda^l_{\gamma_{ij}}}} \tag{4}$$

$$\gamma^l_{ij} = \frac{e^{\lambda^l_{\gamma_{ij}}}}{e^{\lambda^l_{\alpha_{ij}}} + e^{\lambda^l_{\beta_{ij}}} + e^{\lambda^l_{\gamma_{ij}}}} \tag{5}$$

To improve small-target detection, the ASFF multi-scale fusion mechanism is added to the detection layer (Figure 5(d)). It performs spatially weighted fusion across levels and allows dynamic adjustment of anchor boxes and stride during inference.

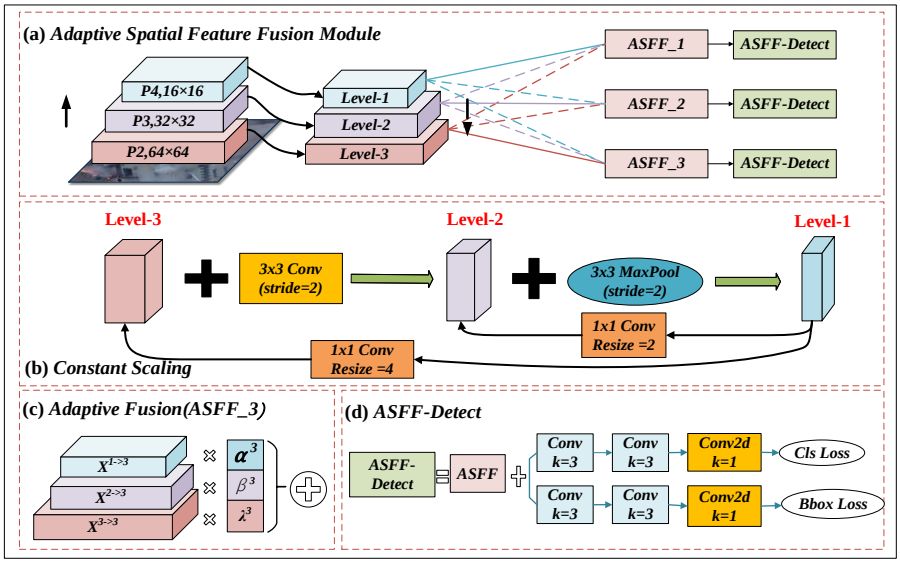


Figure 5. ASFF-Detect module structure design

(5) Miniature Detection Structure

YOLO11 uses three detection heads to handle objects of different sizes. However, small aerial targets lose information after multiple downsampling steps, reducing detection performance. As shown in Figure 6, the proposed model removes the large-scale heads and adds an upsampling structure with a Micro-Head to enhance small-object perception and improve accuracy while reducing parameters and computation.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202512_11(12).0007

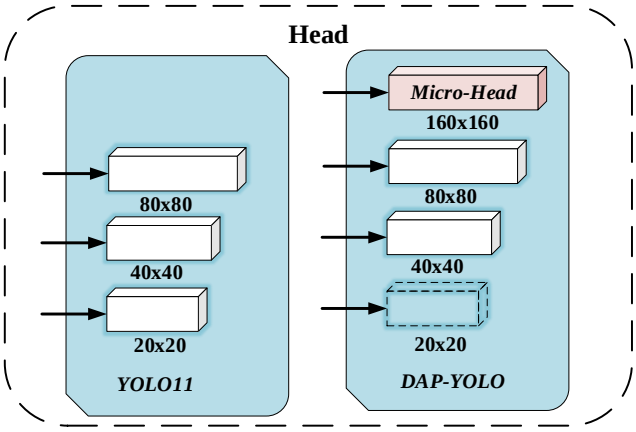


Figure 6. Micro-Head structure

3.3. Experimental Configuration and Training Strategy Design

(1) Experimental Environment Configuration

Experiments are conducted on a deep learning server with sufficient computational resources and storage. YOLO11 is implemented in PyTorch with Python 3.10 and GPU acceleration. Table 1 details the environment and software versions to ensure reproducibility and accuracy.

Table 1. Configure the experimental environment

Name	Configure
Operating System	Windows11
Development Environment	CUDA 12.4
GPU	NVIDIA GeForce GTX 4080
Deep Learning Architecture	Python-3.10.14, torch-2.4.1

(2) Training Strategy

Optimal parameters were selected via comparative validation (Table 2). Training used 200 iterations, batch size 16, initial learning rate 0.05, input size 640×640, and momentum 0.937. Mosaic data augmentation and a weight decay of 0.0005 were applied to prevent overfitting. The model was optimized using SGD.

Table 2. Key hyperparameters for YOLO11 training

Hyperparameter	Value
Learning Rate	0.05
Batch Size	16
image size	640x640
Momentum	0.937
Weight Decay	0.0005
Epochs	200
Optimizer	SGD

3.4. Datasets and Evaluation Metrics

(1) Dataset Overview

The C2A dataset [22] combines subsets of AIDER and LSP/MPII-MPHB. AIDER provides disaster backgrounds for fire, flood, collapsed structures, and traffic accidents. LSP/MPII-MPHB contains 29,732 human instances in varied poses, including bending, kneeling, sitting, standing, and lying. Humans are composed onto disaster scenes using background removal, cropping,

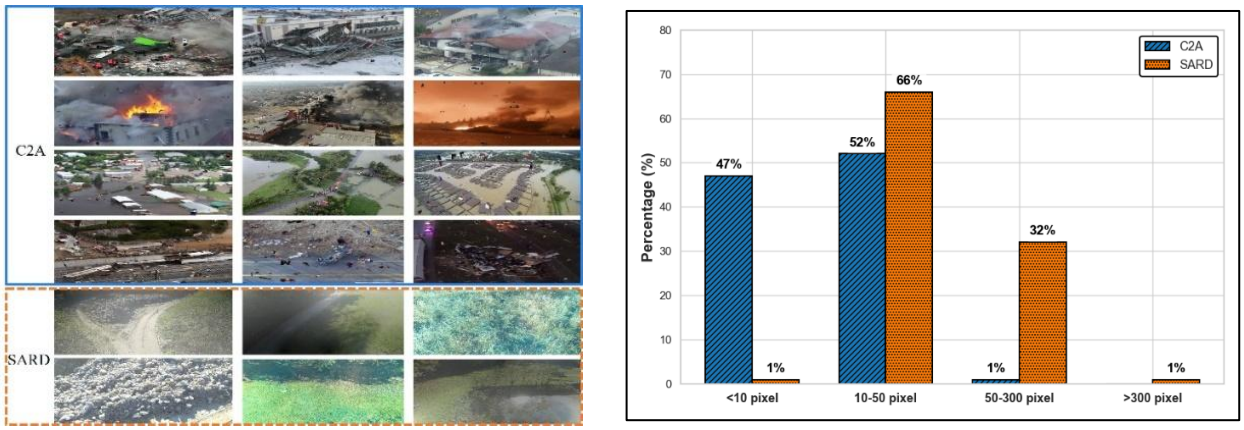
ISSN: 2472-3703 DOI: 10.6911/WSRJ.202512_11(12).0007

scaling, and overlaying. Geometric transformations produce a dataset simulating human detection in disasters. As shown in Figure 7(a), it contains 10,215 images with over 360,000 annotated humans, split 6:2:2 into training, validation, and test sets. Image resolutions range from 123×152 to 5184×3456 pixels.

The SARD dataset [23] contains 3,029 images of humans in diverse injury poses against complex non-urban backgrounds, such as rock piles, grasslands, lakeshores, muddy roads, and forests (Figure 7(a)). Data augmentation includes cropping and random scaling. Various weather and lighting conditions—shadows, haze, heavy rain, blizzards, and color noise—are simulated, with local exposure adjustments to mimic uneven lighting, enhancing image complexity. This realistically simulates disaster rescue scenarios for thorough model evaluation.

As shown in Figure 7(b), most targets in the C2A and SARD datasets are smaller than 50 pixels. Specifically, 47% and 1% are below 10 pixels, while 52% and 66% fall within 10–50 pixels. Few instances exceed 300 pixels. Since small objects are typically defined as occupying less than 0.58% of the image [24], both datasets mainly focus on small-object detection, highlighting the challenges of human detection in disaster scenes.

In summary, the C2A dataset features complex backgrounds and numerous small, occluded human targets, posing major challenges for detection algorithms. It is crucial for real-time and accurate detection in disaster rescue. Therefore, this study uses C2A for algorithm development and SARD for generalization evaluation.



(a) (b)

Figure 7. (a) Data visualization details; (b) Percentage of target instances

3.5. Evaluation Metrics

To assess the detection performance of the proposed improved model, experiments employed precision, recall, mAP50, mAP50-95, number of parameters, model size, inference time, and F1-score as evaluation metrics.

Precision is the ratio of correctly predicted positives to all predicted positives, while recall is the ratio of correctly predicted positives to all actual positives. These metrics help minimize false positives and false negatives, which are vital for accurate human detection in disaster rescue. TP, FP and FN represent correctly detected, falsely detected, and missed targets, respectively. Calculations follow Equations (6-7):

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202512_11(12).0007

Average Precision (AP) measures overall detection accuracy by combining precision and recall across different thresholds, as shown in Equation (8).

$$AP = \int_0^1 P(R)dR \tag{8}$$

Mean Average Precision (mAP) represents the weighted average of AP across all classes. This study mainly uses mAP50 and mAP50–95, the latter providing a stricter measure of localization and classification accuracy, as shown in Equation (9).

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{9}$$

F1-score evaluates classification performance by combining precision and recall, making it suitable for imbalanced datasets, as shown in Equation (10).

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{10}$$

Model size refers to the storage space (MB). The number of parameters indicates how many learnable weights are optimized during training. Inference time shows how long the model takes to make predictions(ms), which is crucial for real-time or limited-resource tasks.

4. EXPERIMENTAL RESULTS

4.1. Hyperparameter Optimization Experiments

To find optimal hyperparameters, experiments tested combinations of batch size (4, 8, 16, 32), optimizer (Adamax, Adam, Nadam, SGD), and initial learning rate (0.1, 0.05, 0.01, 0.005). A grouped tuning strategy optimized one parameter at a time, keeping the others at their best values. Results (Table 3) show: batch size 16 outperformed 4, 8, and 32, as very small or large batches reduce stability and generalization; SGD was the most robust optimizer under complex conditions; a learning rate of 0.05 gave the best F1-score and fastest convergence.

Table 3. Experimental Results under Different Hyperparameters (validation data values)

Hyperparameters		Precision	Recall	mAP50	mAP75	mAP50-95	F1-Score
Batch Size	4	0.844	0.717	0.785	0.556	0.522	77.53%
	8	0.849	0.718	0.788	0.560	0.523	77.8%
	16*	0.849	0.721	0.788	0.560	0.525	77.98%
	32	0.846	0.721	0.788	0.558	0.524	77.85%
Optimizer	Adamax	0.839	0.713	0.779	0.546	0.514	77.09%
	Adam	0.834	0.702	0.767	0.555	0.5	76.23%
	Nadam	0.837	0.708	0.774	0.542	0.51	76.71%
	SGD*	0.849	0.721	0.788	0.560	0.525	77.98%
lr0	0.1	0.844	0.718	0.781	0.554	0.519	77.59%
	0.05*	0.849	0.721	0.788	0.560	0.525	77.98%
	0.01	0.849	0.719	0.787	0.559	0.524	77.86%
	0.005	0.845	0.72	0.785	0.555	0.520	77.75%

Figure 8 shows training results during hyperparameter tuning. “Post Optimisation” refers to the model trained with the final optimal hyperparameters. The study compared learning rates,

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202512_11(12).0007

batch sizes, and optimizers while tracking precision, recall, mAP50, and mAP50-95. Batch size 32 slightly improved mAP50, learning rate 0.005 with batch 32 gave recall similar to the optimized model. Overall, the "Post Optimisation" model was more stable and outperformed others in mAP50, mAP50-95, recall, and F1-score, showing superior detection and convergence.

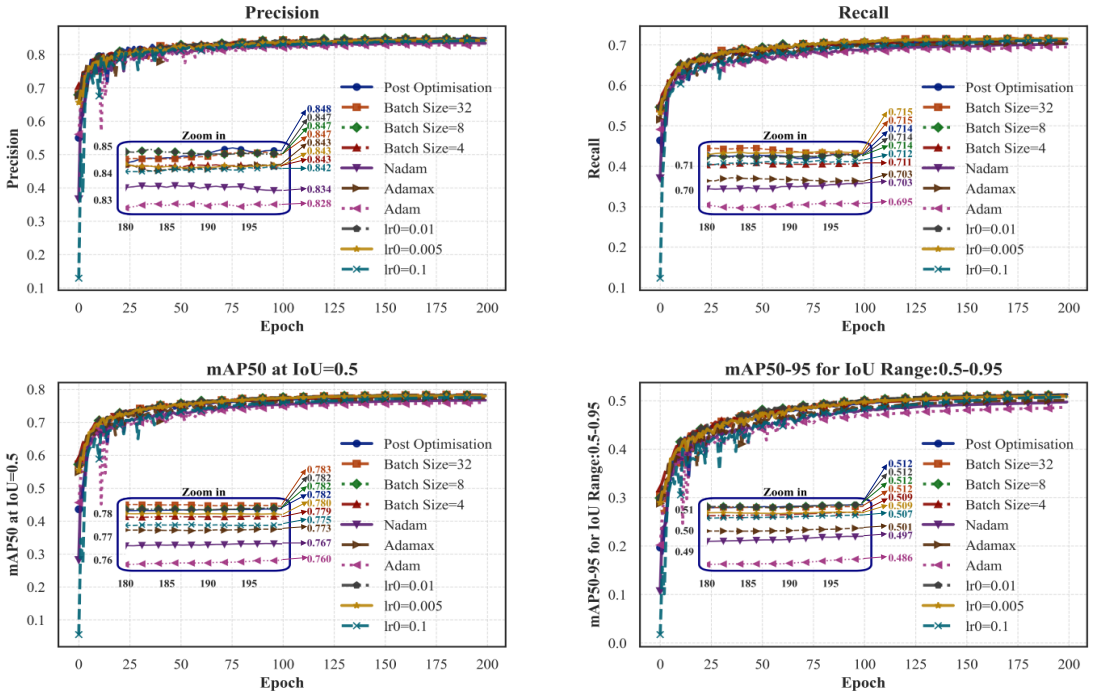


Figure 8. Experimental visualization results for different hyperparameters (training process values)

4.2. Comparative Experiment

Table 4. Classical model comparison experiment

Models	mAP50	mAP75	mAP50-95	Model Size	F1-Score
Faster R- CNN	0.622	0.453	0.411	319	53.22%
RetinaNet	0.772	0.484	0.474	279	65.18%
Cascade R-CNN	0.635	0.470	0.430	536	54.74%
CenterNet	0.739	0.420	0.424	248	62.20%
ATSS	0.541	0.243	0.272	248	44.30%
Dino	0.697	0.361	0.377	565	57.33%
YOLOX	0.819	0.561	0.530	119	69.86%
RT-DETR	0.709	0.400	0.399	64.6	73.13%
RTMDet-tiny	0.637	0.354	0.357	77.7	53.29%
DAP-YOLO	0.841	0.627	0.583	11.0	82.61%

As shown in the comparative results of classical models in Table 4, Faster R-CNN struggles with small objects due to downsampling. RetinaNet mitigates class imbalance with Focal Loss but adapts poorly to complex scenes. Cascade R-CNN performs well but is computationally heavy. CenterNet is lightweight but fails in dense or occluded cases. ATSS adapts samples but struggles with small objects in complex backgrounds. DINO improves detection with denoising and dynamic anchors but is computationally costly. YOLOX improves small and dense object detection but remains complex. RT-DETR and RTMDet-tiny balance speed and accuracy but have limited small object performance. Overall, DAP-YOLO outperforms these models across metrics while keeping parameters low, achieving accurate and efficient small object detection.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202512_11(12).0007

To evaluate DAP-YOLO in disaster human detection, it was compared with lightweight YOLO models and recent baselines (Table 5). Compared to the baseline model YOLO11n, DAP-YOLO achieves improvements of 5.3%, 6.7%, and 5.8% in mAP50, mAP75, and mAP50–95, respectively. When compared to YOLOv12n, DAP-YOLO demonstrates gains of 5.4%, 5.9%, and 5.1%. YOLOv13n is smaller but less accurate; YOLOv5n has slightly better F1 but poor small-object detection. YOLOv6, YOLOv8n, and YOLOv10n are faster or smaller but show lower accuracy. YOLOv8-world is larger yet less accurate. Overall, DAP-YOLO achieves the best balance of accuracy, size, and speed, showing strong robustness and adaptability for small, occluded, and complex disaster scenarios.

Table 5. YOLO series comparison experiment

Models	mAP50	mAP75	mAP50-95	Inference time	Model size	F1-Score
YOLOv5l	0.832	0.569	0.534	3.5	13.7	83.14%
YOLOv6	0.763	0.539	0.507	3.8	8.60	76.23%
YOLOv8n	0.786	0.556	0.522	2.9	5.37	77.79%
YOLOv8-world	0.782	0.561	0.526	3.0	15.8	77.78%
YOLOv10n	0.787	0.560	0.524	3.2	5.70	77.49%
YOLO11n	0.788	0.560	0.525	2.8	5.50	77.98%
YOLOv12n	0.787	0.568	0.532	3.2	5.50	78.04%
YOLOv13n	0.773	0.533	0.504	3.5	5.40	76.91%
DAP-YOLO	0.841	0.627	0.583	3.8	11.0	82.61%

Figure 9 compares mAP50–95 and model size for YOLO and classic detectors. Two-stage detectors are usually larger than 200MB but show limited accuracy gains, indicating high computational cost. YOLO models achieve high accuracy with sizes mostly within 0–50MB, showing a lightweight advantage. DAP-YOLO further improves high-threshold mAP50–95 while keeping a small model size, balancing accuracy and efficiency.

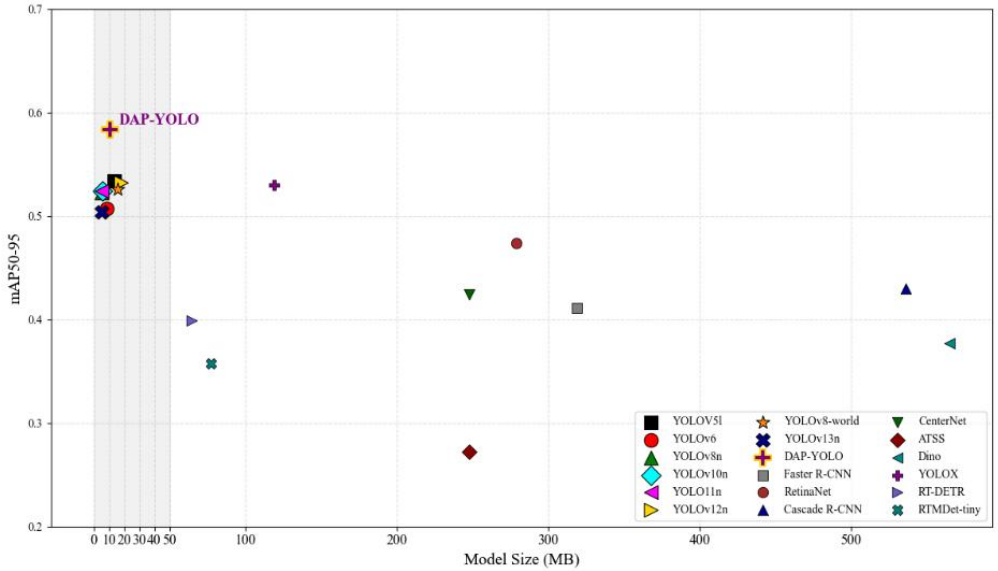


Figure 9. Visualization of the detection accuracy and model size of the classic models with the YOLO series

4.3. Ablation Studies

To systematically evaluate the effectiveness of the proposed architectural improvements for small object detection, two sets of ablation studies were conducted on the C2A dataset using the baseline model YOLO11n. These studies involved comparative performance analyses

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202512_11(12).0007

focusing on detector head adjustments and module integration optimization. Five models tested different detector heads: baseline, 160×160 (Model 1), 160×160 without 20×20 (Model 2), 320×320 (Model 3), and 320×320 without 20×20 (Model 4). Table 6 shows Model 2 achieved the best balance with highest precision (85.3%), recall (85.3%), mAP50 (82.7%), mAP50–95 (54.9%), and F1-score (81.12%), outperforming others in overall detection efficiency. Model 2 was thus chosen as the final detector head, maximizing small object detection while remaining lightweight.

Table 6. Detection of head ablation experiments

Models	Precision/%	Recall/%	mAP50/%	mAP50-95/%	Params/10^5	Inference time	F1-Score
Base	84.9	72.1	78.8	52.5	25.89	2.8	77.98%
Model 1	85.2	76.8	82.1	54.2	26.66	3.5	80.81%
Model 2	85.3	77.3	82.7	54.9	19.24	3.3	81.12%
Model 3	84.8	76.8	81.7	53.9	26.48	3.3	80.64%
Model 4	85.2	77.0	82.3	54.4	19.19	3.4	80.90%

To visually demonstrate the effectiveness of each added or improved module, the following ablation experiment design was devised:

A: Base model architecture.

B: Enhanced C3k2_DWR module.

C: Adaptive feature fusion extraction detection branch (ASFF-Detect).

D: Model 2 detection architecture.

E: Introducing the PPA module.

Table 7 shows the impact of each module in DAP-YOLO on the C2A dataset. Replacing C3k2 with C3k2_DWR improved precision and mAP50–95. Adding ASFF-Detect further enhanced all metrics. Incorporating Model 2 and PPA increased mAP50–95 by 3.0% and 2.0%, respectively. Overall, integrating all modules boosted recall by 6.7% and mAP50, mAP50–95, F1 by 5.3%, 5.1%, and 4.63%. These results confirm that the multi-module optimization improves small object detection, reduces false negatives, and maintains a lightweight design.

Table 7. Improve model ablation experiments

A	B	C	D	E	Precision	Recall	mAP50	mAP50-95	Inference Time	F1-Score
YOLO11n					0.849	0.721	0.788	0.525	2.8	77.98%
	√				0.852	0.718	0.787	0.528	2.9	77.93%
		√			0.854	0.725	0.792	0.533	3.0	78.42%
			√		0.859	0.773	0.827	0.555	3.3	81.37%
				√	0.862	0.728	0.797	0.545	3.5	79.03%
	√	√			0.853	0.725	0.790	0.539	3.4	78.38%
	√		√		0.860	0.771	0.828	0.560	3.3	81.31%
	√			√	0.854	0.734	0.798	0.548	3.6	78.92%
		√	√		0.862	0.776	0.833	0.565	3.5	81.67%
		√		√	0.856	0.732	0.798	0.549	3.5	78.92%
			√	√	0.866	0.786	0.839	0.577	3.3	82.41%
	√	√	√		0.859	0.779	0.834	0.569	3.2	81.70%
	√	√	√	√	0.868	0.788	0.841	0.583	3.8	82.61%

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202512_11(12).0007

As shown in Figure 10, after sequentially incorporating each improvement module, the mAP50 and mAP50-95 metrics demonstrated varying degrees of improvement compared to the baseline model. Model 2 notably enhanced small object detection. F1-Score gains indicate a good balance between precision and recall. Together, these improvements help DAP-YOLO reduce missed detections for dense, overlapping, and occlude humans, significantly boosting small object detection performance.

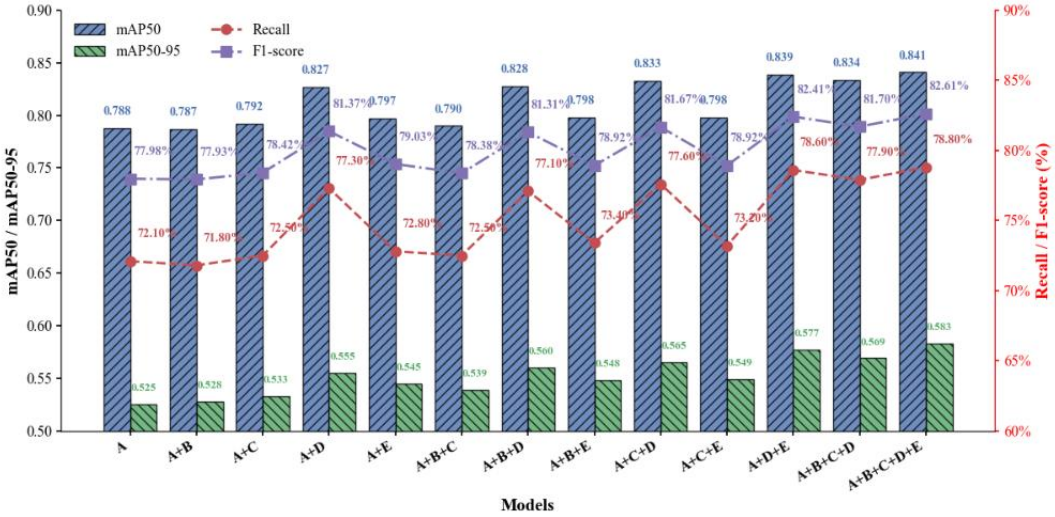


Figure 10. Visualization of ablation experiment results

4.4. Generalization Experiment

To assess generalization in disaster scenarios, DAP-YOLO was evaluated on the SARD dataset for aerial human detection. Compared with YOLO11n (Table 8), it improved precision by 2.9% and recall, mAP50, mAP50–95, and F1-Score by 6.4%, 4.1%, 8.2%, and 4.78%, respectively. These results show DAP-YOLO effectively captures targets under complex lighting, dense scenes, and heavy occlusion, demonstrating strong robustness and practical value for real-world disaster rescue.

Table 8. Detection performance comparison on the SARD dataset

Models	Precision	Recall	mAP50	mAP50-95	F1-Score
YOLO11n	0.935	0.855	0.918	0.582	89.32%
DAP-YOLO	0.964	0.919	0.959	0.664	94.10%

4.5. Visual Analysis

The detection performance of YOLO11n and DAP-YOLO was systematically evaluated on the C2A dataset (Figure 11) using precision, recall, mAP50, and mAP50–95. During the initial 15 training iterations, both models exhibited comparable metric values. As training progressed, DAP-YOLO consistently surpassed YOLO11n, particularly in recall, mAP50, and mAP50–95, indicating enhanced convergence stability and greater optimization potential.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202512_11(12).0007

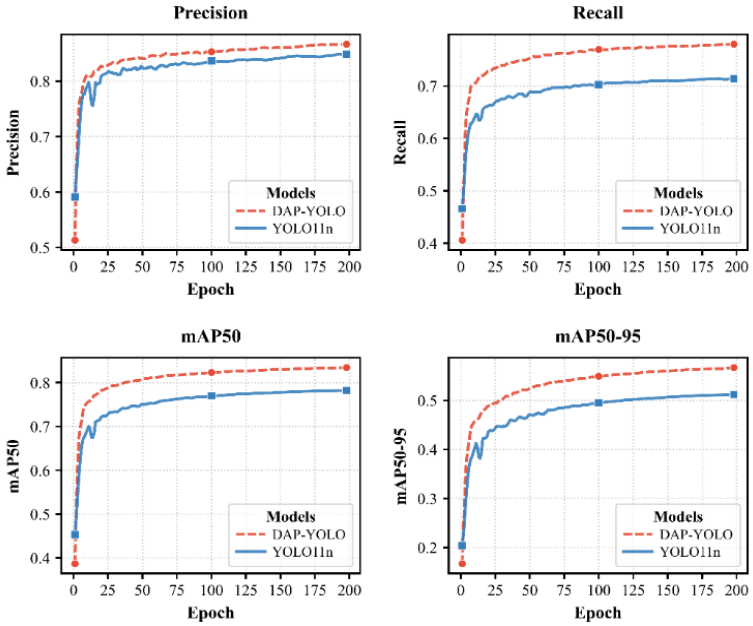


Figure 11. Comparison of evaluation indexes of algorithms before and after improvement

Figure 12 compares YOLO11n and DAP-YOLO detection results across representative disaster scenarios, including building collapses, fires, floods, and nighttime traffic accidents, which involve challenging conditions such as extreme lighting, complex backgrounds, small targets, and dense occlusions. Compared with YOLO11n, which shows notable false positives and negatives, DAP-YOLO accurately detects most previously missed or misclassified targets, including overlapping and non-overlapping small humans (dashed boxes indicate missed or false detections, solid boxes indicate correct detections). These results demonstrate that DAP-YOLO achieves superior robustness and enhanced small object detection in complex disaster environments, effectively reducing false negatives and providing reliable support for search and rescue operations.

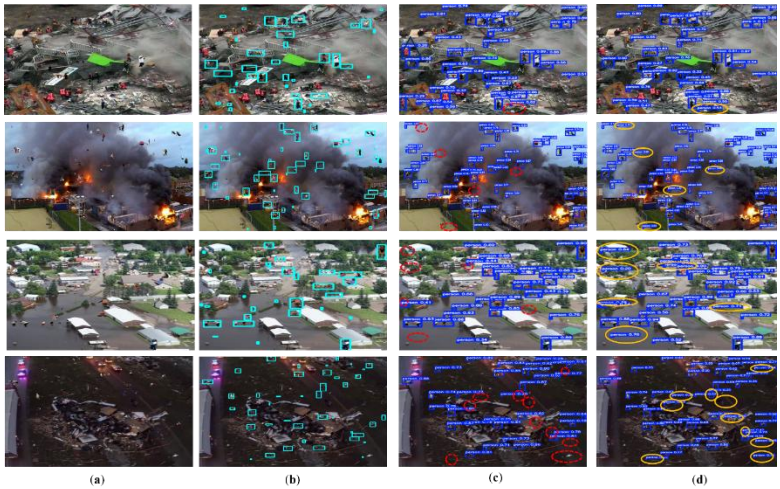


Figure 12. Comparison of detection effectiveness of C2A dataset (a) Original images;(b) Labeling images; (c)YOLO11n ;(d) DAP-YOLO

Heatmaps offer intuitive insights into model attention on critical feature regions and are widely used to assess interpretability. This study applies Grad CAM++[25] to visualize YOLO11n and DAP-YOLO detection results. Grad CAM++ computes gradients between outputs and feature maps to generate spatial attention maps. In Figure 13, red regions indicate high gradient responses, blue regions low. YOLO11n shows predominantly blue or weak responses for small

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202512_11(12).0007

targets, indicating poor attention and frequent misses. DAP-YOLO, in contrast, exhibits stronger, more focused responses in these areas, reducing false negatives and positives. These findings confirm DAP-YOLO's improved attention to small targets and its enhanced detection capability.

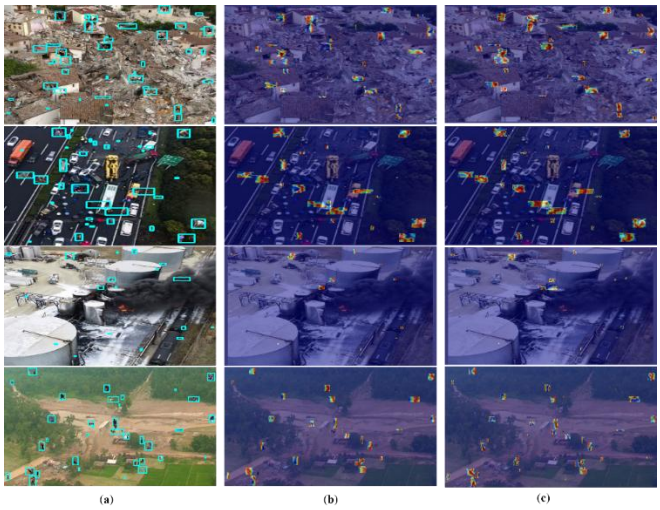


Figure 13. Comparison of heat maps results for C2A dataset (a) Labeling images; (b) YOLO11n; (c) DAP-YOLO

Figure 14 illustrates human detection results of YOLO11n and DAP-YOLO on the SARD dataset. YOLO11n exhibits substantial missed detections and reduced accuracy for small targets in complex disaster scenes. In contrast, DAP-YOLO accurately detects human targets, including multiple objects missed by YOLO11n, demonstrating superior completeness and robustness. This visual analysis confirms DAP-YOLO's effectiveness in handling occlusion, multi-scale targets, and complex backgrounds, highlighting its higher detection accuracy and improved generalization for practical applications.

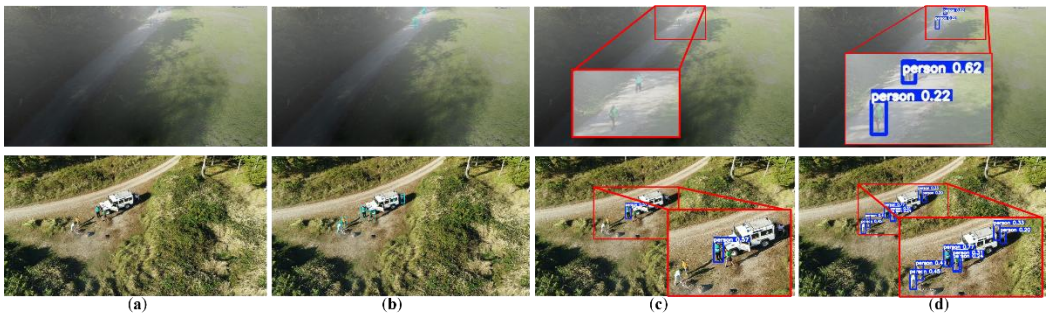


Figure 14. Comparison of detection effectiveness of SARD dataset (a) Original images;(b) Labeling images; (c)YOLO11n ;(d) DAP-YOLO

5. RESULTS

To enhance human detection in disaster scenarios with challenges such as complex backgrounds, small targets, low-resolution images, and severe occlusion, this study proposes DAP-YOLO, a multi-scale adaptive fusion model. The backbone incorporates a PPA module to focus on occluded or variable-pose targets, improving recall and small object recognition. The C3k2_DWR module, based on DWR attention, enlarges the receptive field and integrates multi-scale context for stronger feature extraction. The ASFF-Detect branch enhances scale invariance via adaptive feature fusion, mitigating the effects of dense occlusions and complex environments. A micro-detection architecture tailored for small targets replaces large-scale detection structures and adds an upsampling layer with a micro-detection head, improving accuracy while reducing parameters. Experimental results demonstrate that DAP-YOLO

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202512_11(12).0007

outperforms the baseline model YOLO11n across multiple key metrics: recall on the C2A dataset improves by 6.70%, while mAP50, mAP50-95, and F1-Score increase by 5.30%, 5.10%, and 4.63%, respectively. To evaluate generalization capabilities, the SARD dataset was used for validation. Experiments confirm that DAP-YOLO outperforms YOLO11n in precision, recall, mAP50, and mAP50-95 by 2.90%, 6.40%, 4.1%, and 8.2%, respectively, while achieving an 8.2% improvement in F1-Score. These results validate the effectiveness of DAP-YOLO in small object detection and its robust regulation of false positives and false negatives.

Future work will aim to accelerate inference, reduce computational complexity, and enhance edge deployment. To address small targets, frequent occlusions, and limited data, methods such as self-supervised learning, generative adversarial networks, and few-shot learning will be explored to improve performance under sparse annotations. Additionally, cross-modal fusion and multi-source remote sensing collaborative modeling will be investigated to strengthen generalization and robustness in complex disaster scenarios, providing efficient and reliable technical support for emergency response.

DATA AVAILABILITY STATEMENT

This paper was supported by available online at https://github.com/Ragib-Amin-Nihal/C2A and https://www.kaggle.com/datasets/kushagrgoyal/sard-yolo.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This research was funded by the National Natural Science Foundation of China (U1304402, 41977284) and the Laboratory of Mine Spatio-Temporal Information and Ecological Restoration, MNR (No.KLM202310).

REFERENCES

[1] C. Xu, Z. Xue, "Applications and challenges of artificial intelligence in the field of disaster prevention, reduction, and relief," Natural Hazards Research, 4(1), Vol.(2024), pp. 169-172.

[2] L. Jiang, B. Yuan, J. Du, et al., "MFFSODNet: Multiscale Feature Fusion Small Object Detection Network for UAV Aerial Images," IEEE Transactions on Instrumentation and Measurement, 73, Vol.(2024), pp. 1-14.

[3] X. Luo, Y. Wu, L. Zhao, "YOLOD: A target detection method for UAV aerial imagery," Remote Sensing, 14(14), Vol.(2022), pp. 3240.

[4] R. Varghese, M. S: YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness, 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)* 18-19 April 2024), p. 1-6.

[5] W. Liu, D. Anguelov, D. Erhan, et al.: Ssd: Single shot multibox detector, Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, p. 21-37.

[6] N. Carion, F. Massa, G. Synnaeve, et al.: End-to-end object detection with transformers, European conference on computer vision, p. 213-229.

[7] X. Zhu, W. Su, L. Lu, et al., "Deformable detr: Deformable transformers for end-to-end object detection," arXiv preprint arXiv:2010.04159, Vol.(2020).

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202512_11(12).0007

[8] Z. Liu, Y. Lin, Y. Cao, et al.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, 2021 IEEE/CVF International Conference on Computer Vision (ICCV)* 10-17 Oct. 2021), p. 9992-10002.

[9] W. Wang, E. Xie, X. Li, et al.: Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions, 2021 IEEE/CVF International Conference on Computer Vision (ICCV)* 10-17 Oct. 2021), p. 548-558.

[10] J. Ren, C. Niu,J. Han, "An IF-RCNN Algorithm for Pedestrian Detection in Pedestrian Tunnels," IEEE Access, 8,Vol.(2020), pp. 165335-165343.

[11] J. Zhang, F.-W. Li, W.-Z. Nie, et al., "Visual attribute detction for pedestrian detection," Multimedia Tools Appl., 78(19), Vol.(2019), pp. 26833–26850.

[12] G. Wang, Y. Chen, P. An, et al., "UAV-YOLOv8: A Small-Object-Detection Model Based on Improved YOLOv8 for UAV Aerial Photography Scenarios," Sensors, 23(16), Vol.(2023), pp. 7190.

[13] E. T. Alotaibi, S. S. Alqefari, A. Koubaa, "LSAR: Multi-UAV Collaboration for Search and Rescue Missions," IEEE Access, 7, Vol.(2019), pp. 55817-55832.

[14] C. Qiu, D. Zhang, Y. Hu, et al., "Radio-Assisted Human Detection," Trans. Multi., 25, Vol.(2023), pp. 2613–2623.

[15] H. Zhang, W. Sun, C. Sun, et al., "HSP-YOLOv8: UAV Aerial Photography Small Target Detection Algorithm," Drones, 8(9), Vol.(2024), pp. 453.

[16] B. Du, Y. Huang, J. Chen, et al.: Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, p. 13435-13444.

[17] L. Zhao, M. Zhu, "MS-YOLOv7:YOLOv7 Based on Multi-Scale for Object Detection on UAV Aerial Photography," Drones, 7(3), Vol.(2023), pp. 188.

[18] X. Zhang, Y. Feng, S. Zhang, et al., "Finding Nonrigid Tiny Person With Densely Cropped and Local Attention Object Detector Networks in Low-Altitude Aerial Images," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15, Vol.(2022), pp. 4371-4385.

[19] J. Zhao, W. Yang, F. Wang, et al., "Research on UAV Aided Earthquake Emergency System," IOP Conference Series: Earth and Environmental Science, 610(1), Vol.(2020), pp. 012018.

[20] H. Wei, X. Liu, S. Xu, et al., "DWRSeg: Rethinking efficient acquisition of multi-scale contextual information for real-time semantic segmentation," arXiv preprint arXiv:2212.01173, Vol.(2022), pp.

[21] S. Liu, D. Huang, Y. Wang, "Learning Spatial Fusion for Single-Shot Object Detection," ArXiv, abs/1911.09516, Vol.(2019), pp.

[22] R. A. Nihal, B. Yen, K. Itoyama, et al.: UAV-Enhanced Combination to Application: Comprehensive Analysis and Benchmarking of a Human Detection Dataset for Disaster Scenarios, International Conference on Pattern Recognition, p. 145-162.

[23] S. Sambolek, M. Ivasic-Kos, "Automatic Person Detection in Search and Rescue Operations Using Deep CNN Detectors," IEEE Access, 9, Vol.(2021), pp. 37905-37922.

[24] Q. Tang, C. Su, Y. Tian, et al., "YOLO-SS: optimizing YOLO for enhanced small object detection in remote sensing imagery," The Journal of Supercomputing, 81(1), Vol.(2024), pp. 303.

[25] A. Chattopadhay, A. Sarkar, P. Howlader, et al.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, 2018 IEEE winter conference on applications of computer vision (WACV), p. 839-847.