Article

Not peer-reviewed version

A Lightweight Multi-Scale Context Detail Network for Efficient Military Target Detection in Resource-Constrained Environments

Kaipeng Wang , Guanglin He * , Xinmin Li

Posted Date: 18 April 2025

doi: 10.20944/preprints202504.1498.v1

Keywords: Military target detection; Lightweight neural network; Multi-scale feature fusion; Context-aware modulation; Edge computing

Article

A Lightweight Multi-Scale Context Detail Network for Efficient Target Detection in Resource-Constrained Environments

Kaipeng Wang, Guanglin He * and Xinmin Li

Science and Technology on Electromechanical Dynamic Control Laboratory, Beijing Institute of Technology, Beijing 100081, China; 3120215105@bit.edu.cn (K.W.);3120225110@bit.edu.cn (X.L.)

* Correspondence: heguanglin@bit.edu.cn;Tel.:+86-150-1096-5661

Abstract: Target detection in resource-constrained environments faces multiple challenges, such as the use of camouflage, diverse target sizes, and harsh environmental conditions. Moreover, the need for solutions suitable for edge computing environments, which have limited computational resources, adds complexity to the task. To meet these challenges, we propose MSCDNet (Multi-Scale Context Detail Network), an innovative and lightweight architecture designed specifically for efficient target detection in such environments. MSCDNet integrates three key components: the Multi-Scale Fusion Module, which improves the representation of features at various target scales; the Context Merge Module, which enables adaptive feature integration across scales to handle a wide range of target conditions; and the Detail Enhance Module, which emphasizes preserving crucial edge and texture details for detecting camouflaged targets. Extensive evaluations highlight the effectiveness of MSCDNet, which achieves 40.1% mAP50-95, 86.1% precision, and 68.1% recall, while maintaining a low computational load with only 2.22M parameters and 6.0G FLOPs. When compared to other models, MSCDNet outperforms YOLO-family variants by 1.9% in mAP50-95 and uses 14% fewer parameters. Additional generalization tests on VisDrone2019 and BDD100K further validate its robustness, with improvements of 1.1% in mAP50 on VisDrone and 1.2% in mAP50-95 on BDD100K over baseline models. These results affirm that MSCDNet is well-suited for tactical deployment in scenarios with limited computational resources, where reliable target detection is paramount.

Keywords: Target detection; Lightweight neural network; Multi-scale feature fusion; Context-aware modulation; Edge computing

1. Introduction

Target detection in resource-constrained environments is a major challenge. By studying target detection algorithms in resource-constrained environments, we can provide great help for personnel search and rescue, traffic management and other civilian fields. As the representative of resource-constrained environments, the battlefield environment includes desert, jungle, mountain, city, ocean, etc. The diversified environment can provide a good research support for us. By studying the military target recognition algorithm in the battlefield environment, its technology can be further applied to personnel search and rescue, traffic management and other aspects. The battlefield environment is a complicated, informatization data space， and the detection algorithms must be most incredibly responsive and specific, adapting to a continuously changing panorama for more and more intelligence accumulating and opportunity evaluation [1-3]. The development of efficient and robust target detection strategies has become a hot research direction in resource-constrained environments.

The target detection algorithm in resource-constrained environments is constantly optimized. The first such approaches were based on segmentation techniques, which have separated possible targets from the background elements [4]. Fuzzy inference systems were later added to these to deal

doi:10.20944/preprints202504.1498.v1

with environmental uncertainties and classification ambiguities [5]. Breakthroughs further took the form of Charge Coupled Device image processing for greater detection in variable lighting [6]. All weather detection of submarines was obtained by fusing SAR data to early detection schemes, greatly improving capabilities [7,8].

Being applied as a target detection milestone, the ICA was a key milestone in the evolution of military target detection. Idea of ICA is shown by Twari et al. [9] in which it is used on hyperspectral images to detect military targets, which provides better distinction between camouflaged objects and their surrounding environment. Consequently, spectrally modeled algorithms in conjunction with ICA resulted in improvements of detection accuracy in a resource-constrained environment that is complex.

However, deployment of such advancements in cluttered, unpredictable operational settings turned out to be a challenge for traditional methods. With high sensitivity to the variations in target appearance and orientation together with the environmental factors, they were dependent on the manually engineered features and thus could not be practically utilized in real resource-constrained scenarios.

The emergence of deep learning to know has revolutionized object detection methodologies, heralding a paradigm shift. The improvement of convolutional neural networks, like SqueezeNet [10], set the level for more effective detection frameworks. The following improvement of the area-based totally CNN (R-CNN) family, such as rapid R-CNN [11], Faster R-CNN [12], and Mask R-CNN [13], resolved some of the shortcomings from the early designs and made the manner for extra robust detection structures.

At the same time, single stage detection frameworks and the Unmarried Shot MultiBox Detector (SSD) [14] as well as the YOLO family of detectors [15] appeared, simultaneously offering computationally efficient solutions that favored speedy detection at the expense of precision. The assignment of finding a balance between computational efficiency and precision, however, keeps forming research in the field.

Target detection in resource-constrained environments is extremely challenging with respect to conventional object detection. But objectives in resource-constrained environments are typically small on a sensor scale of view, are often partially obscured. Also, complex situations with changing climate, not to mention limited computational resources and need for real-time processing, make it intractable to develop powerful detection systems.

Despite such progress in civilian object detection using deep learning frameworks, their use in resource-constrained environments would be constrained. However, although a variety of conventional detection algorithms have been developed using general purpose approaches for general use, the performance of these algorithms in complex environments is not ideal. Two challenges remain: (1) Algorithms need to adapt to complex and diverse environments; (2) The computational burden of the most advanced model exceeds the available resources for tactical problems, which brings great inconvenience to use.

This paper presents a novel approach to target detection in resource-constrained environments that specifically addresses these challenges through a hierarchical feature fusion architecture optimized for multiscale, camouflaged target detection while maintaining computational efficiency. Our key contributions include:

1. A lightweight MSCDNet (Multi-Scale Context Detail Network) architecture that solves the computational resource constraints in resource-constrained environments.
2. A Multi-Scale Fusion (MSF) module that addresses the challenge of detecting targets with significant dimensional variations, camouflage, and partial occlusion.
3. A Context Merge Module (CMM) that overcomes the difficulty of integrating features from different scales for comprehensive target representation.
4. A Detail Enhance Module (DEM) that preserves critical edge and texture details essential for distinguishing camouflaged targets in complex environments.

The remainder of this paper is organized as follows: Section 2 reviews related work in object detection and military target recognition, so as to provide reference for us to solve target recognition under limited resource environments in the civil field. Section 3 details our proposed methodology, including the YOLO11n architecture, Multi-Scale Fusion module, Context Merge Module, and Detail Enhance Module. Section 4 presents experimental results and comparative analysis. Section 5 concludes with a summary of findings and directions for future research.

2. Related work

2.1. Traditional Military Target Detection Methods

Early military target detection relied on distinctive visual characteristics through edge and contour features. Sun et al. [16] proposed adaptive boosting for SAR automatic target recognition, demonstrating how ensemble learning could improve feature-based detection in radar imagery. The application of texture-based features such as Histogram of Oriented Gradients (HOG) [17] and Local Binary Patterns (LBP) improved discrimination between military targets and background elements. Zhang et al. [18] developed face detection based on multi-block LBP representation, a technique that was later adapted for military personnel detection. For non-rigid targets like soldiers, frameworks based on Deformable Part Models (DPM) offered flexibility in handling diverse postures and equipment configurations.

In sensor technology, Pei et al. [19] further explored multiview SAR automatic target recognition optimization, demonstrating how multiple perspectives can enhance detection reliability. Che et al. [20] developed multi-spectral fusion techniques combining infrared and visible light images. Addressing illumination variation issues, while Li et al. [21] specifically researched target classification in low-light night vision conditions.

Motion characteristics of military targets also serve as important cues in traditional methods. Salmon et al. [22] studied the effects of motion on in-vehicle system operation in battle management systems, highlighting practical usability challenges in tactical environments. Bajracharya et al. [23] developed a fast stereo-based system for detecting and tracking pedestrians from moving vehicles, techniques later adapted for soldier tracking. Chaves [24] investigated Kalman filtering for low-cost GPS-based collision warning systems in vehicle convoys, demonstrating practical tracking applications.

Despite achieving success in specific controlled environments, these traditional methods faced numerous challenges: sensitivity to environmental changes, insufficient robustness against camouflaged targets, dependence on expert experience for feature engineering, and lack of end-to-end learning capabilities. Boult et al. [25] addressed some of these limitations through specialized visual surveillance systems for non-cooperative and camouflaged targets in complex settings, but fundamental challenges remained that motivated researchers to transition toward deep learning methods.

2.2. General Deep Learning Methods for Military Target Detection

The emergence of deep learning, particularly Convolutional Neural Networks (CNNs), has fundamentally transformed military target detection research. Zhao et al. [26] applied the Faster R-CNN framework to storage tank detection using high-resolution aerial imagery, significantly improving detection accuracy. Peng et al. [27] utilized Faster R-CNN for multi-object extraction in complex backgrounds under artificial intelligence contexts, addressing background interference issues. Naz et al. [28] explored soldier detection using unattended acoustic and seismic sensors, optimizing detection systems for various battlefield conditions.

Single-stage detectors such as YOLO and SSD have also been widely applied to military scenarios due to their efficiency. Xu and Wu [29] improved the YOLOv3 model with DenseNet for multi-scale remote sensing target detection, achieving balance between speed and accuracy. Wang et al. [30] developed a lightweight detector based on SSD with depth-separable convolution, balancing

doi:10.20944/preprints202504.1498.v1

computational efficiency and detection performance. These general object detection frameworks provide fundamental solutions for military target detection but require domain-specific adaptations to address unique military challenges.

With advancements in deep learning technology, emerging network architectures have been continuously applied to military target detection. Feature Pyramid Networks (FPN) have been widely used to address scale variations in military targets. Wang et al. [31] developed a multi-scale infrared military target detection system based on 3X-FPN feature fusion network, capable of processing targets at different scales. RetinaNet and its Focal Loss have been applied to address the foreground-background class imbalance in detection tasks. Liu et al. [32] improved RetinaNet for high precision detection of transmission line defects, effectively enhancing detection sensitivity.

Transformer architecture has recently been introduced to military target detection with promising results. Pushkarenko and Zaslavskyi [33] researched areas in Ukraine affected by military actions using remote sensing data and deep learning architectures. Li et al. [34] developed a military target detection framework based on Swin Transformer, utilizing SwinF with feature fusion for enhanced target detection. These advanced architectures demonstrate superior performance in modeling long-range dependencies and contextual relationships, crucial for understanding complex battlefield scenarios.

To address specific requirements of military applications, researchers have developed specialized techniques. Zhuang et al. [35] proposed military target detection methods based on EfficientDet and Generative Adversarial Networks, simulating conditions such as smoke, dust, and partial occlusion. Sun et al. [36] introduced YOLO-E, a lightweight object detection algorithm specifically designed for military targets. Jani et al. [37] reviewed model compression methods for YOLOv5, addressing deployment constraints in resource-limited tactical environments. Zhang et al. [38] investigated multi-scale feature fusion networks for object detection in very high resolution optical remote sensing images, lowering deployment thresholds for field operations.

2.3. Deep Learning Methods for Specific Target Detection

The deep learning method can solve the unique characteristics of different targets, and take tanks and soldiers as examples for analysis.

For tank detection, Fan et al. [39] developed a fast detection and reconstruction method for tank barrels based on component prior and deep neural networks in the terahertz regime, demonstrating how component-level detection can substantially improve performance. Ma et al. [40] proposed an end-to-end method with transformers for 3-D detection of oil tanks from single SAR images, not only detecting overall position but also identifying key components, providing richer feature information for model recognition.

Addressing camouflage issues in target detection, Song et al. [41] developed a multi-granularity context perception network for open set recognition of camouflaged objects, incorporating spatial and channel attention modules while utilizing contextual information to determine actual target contours against complex terrain backgrounds. Naeem et al. [42] explored multi-sensor fusion technologies, combining data with integrated multi-sensor data fusion algorithms that dynamically adjust sensor importance according to environmental conditions. Lv et al. [43] researched recognition of deformation military targets in complex scenes via MiniSAR submeter images, highlighting the importance of integrating different sensor modalities.

Soldier detection presents different challenges due to non-rigid characteristics, multiple postures, and group behaviors. Wu and Zhou [44] designed video-based martial arts combat action recognition and position detection using deep learning, incorporating modules capable of adapting to morphological changes in different tactical postures. For camouflage detection, Choudhary [45] developed real-time pixelated camouflage texture generation techniques combining multi-scale texture descriptors to capture minute differences between camouflage equipment and natural environments. Barnawi et al. [46] provided a comprehensive review of landmine detection using

doi:10.20944/preprints202504.1498.v1

deep learning techniques, enhancing detection through preprocessing techniques and designing feature extraction networks adapted to challenging environments.

Group behavior analysis for soldiers has also received significant attention. Anzer et al. [47] proposed frameworks incorporating semi-supervised graph neural networks that model spatial relationships for detecting tactical patterns. Wang et al. [48] developed free-walking pedestrian inertial navigation systems based on dual foot-mounted IMU, capable of understanding complex movements. These specialized approaches have significantly advanced target detection technology by addressing the unique challenges of each target type.

2.4. Key Challenges of Target Detection in Resource-constrained Environments

Although target detection based on resource-constrained environments has made progress in deep learning, there are still some challenges. Camouflage and concealment reduce detection accuracy by up to 40%, while scale variation and small target detection are obstacles, especially in aerial reconnaissance where targets occupy less than 1% of the image. Weather conditions like rain, fog, or snow can cause accuracy drops of 35-60%, and while multi-sensor fusion improves all-weather detection, it introduces synchronization and computational challenges. Deployment on edge devices with limited resources creates tension between model compression, energy efficiency, and inference speed. Additionally, limited training data and the complexity of overlapping targets reduce performance by 25-45%, especially in dense formations and dynamic scenarios requiring real-time processing.

To address these issues, we propose the MSCDNet architecture with specialized modules. The Multi-Scale Fusion Module (MSFM) tackles scale variation, the Context Merge Module (CMM) improves feature integration across scales, and the Detail Enhance Module (DEM) preserves critical details for detecting camouflaged or occluded targets. This hierarchical approach balances computational efficiency with enhanced detection performance, addressing the challenges of complex environments.

3. Methodology

3.1. Overview of Model

As illustrated in Figure. 1, the MSCDNet (Multi-Scale Context Detail Network) incorporates a Lightweight Perception Net toward lightweight network of efficient visual feature extract and object detection. The feature representation of this model is constructed gradually across multiple scales yet guaranteed computational efficiency on resource constrained environments. First it is composed of some convolutional layer to set up some basic features and reduce spatial dimensions. C3k2[49] modules with a 2x2 kernel and bottleneck design that process these features are optimized for computational efficiency and representation capacity. The architecture then incorporates three key components for target detection: Multi-Scale Fusion (MSF) modules merge information from diverse receptive fields to capture the multi-scale nature of targets; the Spatial Pyramid Pooling - Fast (SPPF) module aggregates multi-scale contextual information efficiently; and the C2PSA (Cross-Stage Partial with Position-Sensitive Attention) modules refine features with attention mechanisms that highlight relevant information while reducing background noise, improving detection precision.

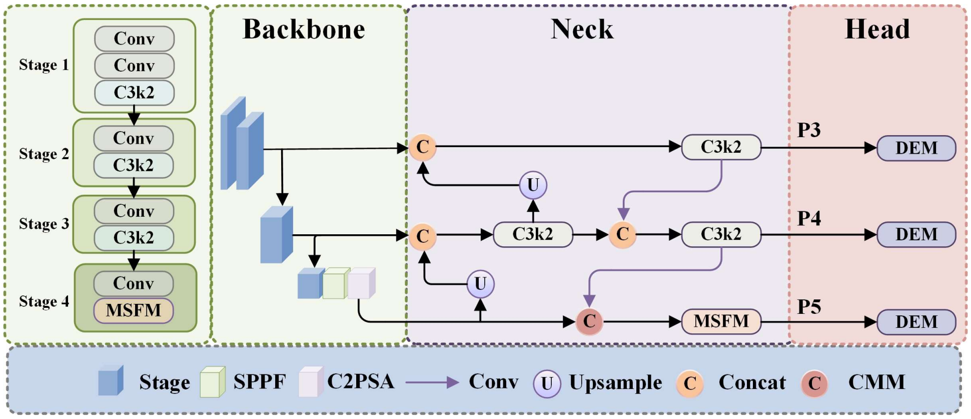


Figure 1. Overall structure diagram.

It implements a design of architecture with an advanced feature pyramid network that combines features across multiple scales through upsampling operations and Concat modules in the detection head. As context aware mechanism and adaptive feature modulation strategies, the Cross-Modal Modulation (CMM) modules replace the simple concatenation approach with a more powerful way of cross-scale feature integration by making use of the complementary information provided by multiple feature levels. C3k2 modules will further amplify the feature's discriminative power after being refined at each scale of the feature pyramid. Finally, these multi scale features pass through the Detail Enhance Module (DEM) that processes them to the final predictions while keeping critical details hence especially important for military target identification.

The perception network obtained in this modular design is a construction, refinement and integration of visual features at different scales. The network is composed of each specialized module, which helps to detect objects of different sizes, and is computationally efficient. Through this seamless coordination, MSCDNet can generate superior detection performance in the presence of complex environments with targets of intricate morphologies, high background interference, and diverse scale, whilst maintaining a lightweight and suitable for constrained resource edge deployments.

3.2. Multi-Scale Fusion Modulation

The use of traditional object detection networks proves difficult when applied to resource-constrained settings. The detection targets include objects of various sizes extending from tiny ones at distance to large equipment near the scene therefore leading to substantial scale variability. The position and shape of the target in the complex environments also have great changes. The detection requires superior algorithms which can extract features efficiently. Basic features extraction techniques succeed during their intended operation though they collapse at capturing multi-scale details and retaining spatial data especially within challenging environments. The MSFM (Multi-Scale Feature Modulation) framework serves as our proposed detection enhancement solution by applying multi-scale convolution strategies and effective feature union approaches in a structure demonstrated in Figure. 2.

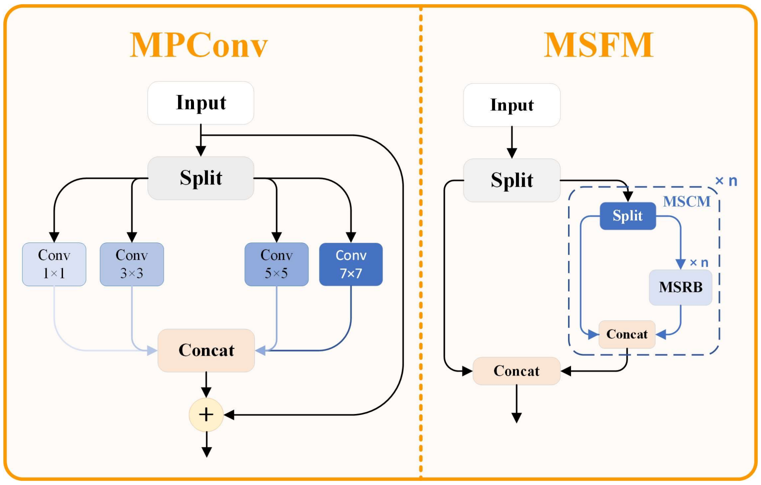


Figure 2. MSFM structure diagram.

The efficient multi-scale convolution modules within the MSFM structure function to improve feature representation quality. Partial feature separation together with advanced multi-scale convolution strategies enables this framework to improve its capability for targets with different scales. The MPConv (Multi-scale Parallel Convolution) module stands as the central aspect of the MSFM by maximizing input information from different receptive fields through parallel multi-scale convolution together with advanced feature reorganization methods.

The MPConv module is the cornerstone innovation of the MSFM structure, and its mathematical formulation is outlined in Equation (1):

$$F_{out} = \phi\left(\sum_{i=1}^{n} \omega_i \cdot (W_i * X_i) + \beta \cdot \sqrt{\prod_{i=1}^{n} \| W_i * X_i \|_F}\right) \cdot \gamma \tag{1}$$

The specific mathematical implementation of the MPConv module can be further decomposed. Initially, the input feature X is uniformly divided into n groups along the channel dimension, with each feature group containing C/n channels when the input feature X has C channels. Subsequently, each feature group is processed through convolution kernels of different sizes, with typical kernel size configurations being $[1 \times 1,3 \times 3,5 \times 5,7 \times 7]$, enabling the model to simultaneously capture spatial information at different scales. The processed features from each group are then reconnected along the channel dimension, and finally, feature fusion and channel dimension adjustment are accomplished through a 1×1 convolution to obtain the final output feature.

The MPConv module functions through a Fourier analysis method that performs multi-scale frequency selection filtering. Organizational structure of convolutional filters includes large and small kernels that extract different frequency ranges where the size of convolutional filters determines the range of information acquired with large filters maintaining contours and big features whereas small ones capture edges and tiny details. The sequence of parallel multi-scale filtering operations maintains plentiful frequency content without eliminating high-frequency details as traditional convolution frameworks would typically do. The frequency response of the filters can be represented via Fourier transformation, as shown in Equation (2):

$$H_i(u,v) = \mathcal{F}(W_i)(u,v) \cdot e^{j\theta_i(u,v)} \cdot \Lambda_i(u,v) \tag{2}$$

where $\mathcal{F}$ denotes the Fourier transform operator, $H_i(u,v)$ represents the frequency response of the i-th convolution kernel, $\theta_i(u,v)$ is the phase response function, and $\Lambda_i(u,v)$ is the filter characteristic curve. The comprehensive module's frequency response is the synthesis of individual components as shown in Equation (3):

$$H(u,v) = \sum_{i=1}^{n} \alpha_i\,(u,v) \cdot H_i(u,v) \cdot \mathcal{F}(X_i)(u,v) + \Omega(u,v) \cdot \sqrt{\prod_{i=1}^{n} |H_i(u,v) \cdot \mathcal{F}(X_i)(u,v)|^2} \tag{3}$$

In the MSFM implementation, the MPConv module is embedded into a basic residual unit, forming a Multi-Scale Residual Block (MSRB) as shown in Equation (4):

$$F_{MSRB} = X + \mathrm{MPConv}\big(\mathrm{Conv}_{1\times1}(X)\big) \tag{4}$$

Here, the initial 1×1 convolution operation compresses the input channels to c (typically half of the output channels), the MPConv operation implements multi-scale feature extraction, and finally, the original features are added to the processed features through a residual connection. This residual connection mechanism not only mitigates the vanishing gradient problem in deep networks but also achieves effective fusion of features at different abstraction levels.

The Multi-Scale Dual-path Module (MSDM) represents a dual-path feature extraction structure that replaces basic units with MSRBs as shown in Equation (5):

$$F_{MSDM} = \mathrm{Concat}\left(X_{path1}, \mathrm{Sequential}\big(\{\mathrm{MSRB}_i\}_{i=1}^{n} \circ \Psi(\theta)\big)\big(X_{path2}\big)\right) \cdot \sqrt{\eta \cdot \mathrm{Norm}\big(X_{path1}\big) \cdot \mathrm{Norm}\big(X_{path2}\big)} \tag{5}$$

In this formulation, the input feature X is divided into two parts, X_{path1} and X_{path2}. The X_{path1} is directly transmitted to the output, while X_{path2} is processed through n serially connected MSRB modules before being transmitted to the output. This dual-path design balances the depth and width of feature extraction, enabling simultaneous preservation of low-level detail features and high-level semantic features.

Ultimately, MSFM employs MSDM as its fundamental building unit as shown in Equation (6):

$$F_{MSFM} = \mathrm{Concat}\left(X_{path1}, \mathrm{ModuleList}(\mathrm{MSDM}_1, \mathrm{MSDM}_2, \ldots, \mathrm{MSDM}_n)\big(X_{path2}\big)\right) \tag{6}$$

This multi-level feature extraction and fusion mechanism can capture rich feature information across various scales, particularly suitable for processing targets with complex morphologies and variable scales in resource-constrained scenarios.

From an information theory perspective, MSFM reduces information loss during feature extraction through parallel multi-scale processing, increasing the mutual information between input features and target features. By maximizing the mutual information as shown in Equation (7):

$$I(X;Y) = H(Y) - H(Y|X) + \Delta(\kappa) \cdot D_{KL}\big(p(Y|X) \parallel q(Y)\big) - \lambda \cdot Tr(\Sigma_{XY}\Sigma_{YX}) \tag{7}$$

where $H(Y)$ represents the entropy of target features, $H(Y|X)$ denotes the conditional entropy of target features given input features, $\Delta(\kappa)$ is an information gain function, D_{KL} is the KL-divergence metric, and $\mathrm{Tr}(\Sigma_{XY}\Sigma_{YX})$ represents the trace operation on feature covariance matrices, MSFM enhances the model's capacity for target recognition and localization with unprecedented accuracy.

Instead of traditional multi-scale feature extraction methods, MSFM has several key advantages. MPConv is a parallel module that processes convolutions at various scales in parallel, thus maintaining the spatial information across scales and keeping away from the information loss in cascaded convolutions occurring with conventional convolutions. The second contribution is that the channel grouping strategy reduces computational complexity, but maintains satisfactory feature extraction capability, and has orders of magnitude lower computational complexity than typical multi-branch structures. Third, MSFM handles the vanishing gradient problem of deep networks when using residual connections for the feature fusion of low- and high-level features. The efficient use of multi-scale convolution in the MSFM structure enables significant improvement of feature representation over multi-scale target identification in resource-constrained detection scenarios.

3.3. Context Merge Module

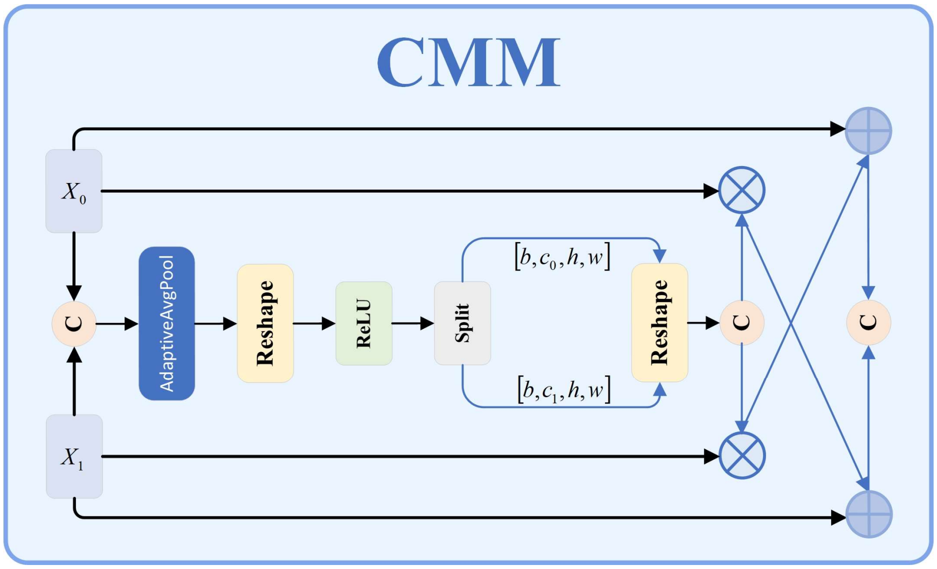


Figure 3. CMM structure diagram.

Traditional object detection networks simply concatenate or fuse the feature by summing the features. Nevertheless, the performance in target detection is poor because many specific targets have complicated shapes, different scales, and strong interference of the background. Conventional fusion techniques do not fully capture complementary information at the different feature levels; hence, they sacrifice efficiency in fusion and target feature capture. The issue of concatenation operations is that they simply add features of different levels without considering contextual correlations or adapting to their own features, which leads to reduction in detection accuracy. We propose the Context Guided Modulation Module (CMM)as shown in Figure 3, which involves context aware mechanisms and adaptive modulation strategies to improve the feature extraction, and to achieve higher accuracy and robustness of specific target recognition.

The key idea of the CMM design is to leverage contextual information to enable the adaptive fusion of features at varying levels to achieve synergy between different levels' features, making the most of the complementary feature relations to improve the model's feature roughness. In the module, there are four key components: feature adjustment layer, feature concatenation operation, channel attention mechanism, and complementary feature fusion. The CMM module first makes a feature adjustment layer to make the channel dimension compatible between features at different levels. When the number of channels in input feature x_0 differs from that in feature x_1, a 1×1 convolution operation is applied to feature x_0 to adjust its channel count to match that of feature x_1.

Second, the adjusted feature x'_0 is concatenated with feature x_1 along the channel dimension to form a feature concatenation vector x_{concat}.

Third, the concatenated feature is processed through a Squeeze-and-Excitation (SE) attention module, which first performs global average pooling on the feature to capture global contextual information between channels, then learns the correlations between channels through a non-linear transformation constructed by two fully connected layers, ultimately generating channel attention weights, as shown in Equation (8):

$$\mathrm{W} = \sigma(\mathrm{W}_2 \cdot \delta(\mathrm{W}_1 \cdot \mathrm{GAP}(\mathrm{x}_{\mathrm{concat}}) + \mathrm{b}_1) + \mathrm{b}_2) \cdot \alpha + \beta \cdot \tanh\big(\gamma \cdot \mathrm{GAP}(\mathrm{x}_{\mathrm{concat}})\big) \tag{8}$$

where GAP denotes the global average pooling operation, W_1 and W_2 represent the weight matrices of the respective fully connected layers, b_1 and b_2 denote the corresponding bias terms, δ signifies the ReLU activation function, σ represents the Sigmoid activation function, while α, β, and γ are learnable scaling parameters, and the tanh function provides additional non-linearity to enhance expressiveness.

doi:10.20944/preprints202504.1498.v1

The generated channel attention weights W are subsequently bifurcated into two segments corresponding to weights for x'_0 and x_1 respectively. These weights are then applied to the original features through element-wise multiplication to derive weighted feature representations.

Finally, through a cross-enhancement feature strategy, the weighted features and original features undergo complementary fusion, as expressed in Equation (9):

$$x_{\text{out}} = \text{Concat}\left(\lambda_1 \cdot \left(x'_0 + \Phi\left(x_{1_{\text{weighted}}}, \theta_1\right)\right) \oplus \lambda_2 \cdot \left(x_1 + \Psi\left(x_{0_{\text{weighted}}}, \theta_2\right)\right)\right) \cdot \sqrt{\eta \cdot \text{Norm}(x'_0) \cdot \text{Norm}(x_1)} \tag{9}$$

where Φ and Ψ represent non-linear transformation functions with parameters θ_1 and θ_2 respectively, λ_1 and λ_2 are balancing factors, $\oplus$ denotes channel-wise concatenation, Norm signifies feature normalization, and η is a global scaling factor.

By preserving the original features and subsequently integrating context enhanced information from other features, the Context guided Modulation Module (CMM) helps the feature fusion remain efficient by preserving complementarity. Within the enhanced YOLO11 network, the CMM is strategically made to replace traditional feature concatenation as part of the multi-scale feature fusion. It is used at critical stages (from higher level to lower-level features, P_5→P_4, P_4→P_3; from lower level (P_3) to higher level (P_4, P_4) features). Optimizing feature data to each scale of the environment, this multi-tiered approach is an improvement to the detection of military targets at multiple sizes. The CMM retains diverse feature characteristics by combining channel attention with complementary fusion, adapting to them to weight and integrate more precisely. Finally, cross enhancement further improves the accuracy of detection of specific targets like small and multi scale targets in cluttered environments. Experiment results demonstrate the advantage of using YOLO11 network over the CMM for specific target detection. CMM's principles also suggest ways of dealing with feature fusion issues in other computer vision areas.

3.4. Detail Enhance Module

The Detail Enhance Module is a step up from object detection technology, mainly as regards to applications such as tanks and soldier detection. Based on strong capabilities in advanced detection frameworks, this detection head incorporates the modifications that help feature representation and detection precision, especially for difficult targets. DEM architecture pulls together grouped normalization and detail enhanced convolutions to bring in an innovative way of feature processing. The innovation resides in the shared convolutional structure, which is trained to capture the details at the boundaries that are imperative for precise identification of specific targets against clutter and challenging environments. The DEM structure diagram is shown in Figure 4.

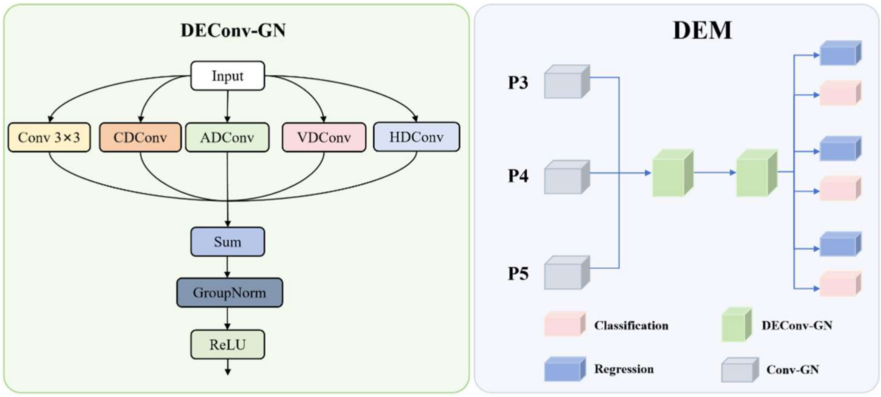


Figure 4. DEM structure diagram

The DEM architecture centers around the Detail Enhanced Convolution (DEConv-GN) module, which enhances edge and texture details. The input splits into five parallel convolution paths: Standard Conv 3×3 for baseline feature extraction, CDConv for emphasizing central pixel variations, HDConv for highlighting horizontal edges, VDConv for vertical edges, and ADConv for capturing diagonal features. These paths combine through a Sum operation, followed by GroupNorm and ReLU activation. In DEM, DEConv-GN replaces batch normalization with group normalization to improve stability across varying batch sizes and conditions typical in surveillance. The detection head processes multi-scale features (P3, P4, P5) through initial Conv-GN layers, followed by a shared detail-enhanced convolution module with two DEConv-GN layers. This module ensures efficient cross-scale feature learning, after which the features are split for regression and classification tasks.

Mathematically, the feature transformation process reduces the input feature map dimensions, effectively compressing channel dimensions while preserving discriminative information necessary for accurate specific target detection. Detail enhancement occurs in the shared module as shown in Equation (10):

$$F_i'' = \mathtt{DEConvGN}_2(\mathtt{DEConvGN}_1(F_i')) \tag{10}$$

The DEConv operations can be collectively expressed as shown in equation (11):

$$\mathrm{DEConv}(x) = \mathrm{Act}\left(\mathrm{Norm}\left(\sum_{j\in\{\mathrm{cd,hd,vd,ad,std}\}} \omega_j \cdot \mathrm{Conv}_j(x) + \gamma \cdot \sqrt{\prod_{j\in\{\mathrm{cd,hd,vd,ad,std}\}} \| \mathrm{Conv}_j(x) \|_2}\right)\right) \tag{11}$$

where ω_j represents adaptive weights for each convolution branch learned during training, γ denotes the cross-branch interaction coefficient, $\| \mathrm{Conv}_j(x) \|_2$ calculates the L2-norm of convolution outputs to represent feature significance, Norm implements the Group normalization function with 32 groups, and Act applies a non-linear activation function such as ReLU or Swish.

For efficient deployment, the DEConv module can be simplified by fusing the parallel convolution branches into a single convolution operation as shown in Equations (12) and (13):

$$W_{\mathrm{fused}} = \sum_{j\in\{\mathrm{cd,hd,vd,ad,std}\}} \lambda_j \cdot W_j \cdot \Phi_j(\Theta_j) \tag{12}$$

$$b_{\mathrm{fused}} = \sum_{j\in\{\mathrm{cd,hd,vd,ad,std}\}} \mu_j \cdot b_j \cdot \Psi_j(\Omega_j) \tag{13}$$

where W_j represents convolution weight matrices for each branch, b_j denotes the bias vectors for each branch, λ_j and μ_j are branch importance coefficients determined during the optimization phase, and $\Phi_j(\Theta_j)$ and $\Psi_j(\Omega_j)$ implement transformation functions with learnable parameters Θ_j and Ω_j that enable adaptive branch fusion during model compression.

Following the shared feature processing, the detection head splits into two parallel branches: a regression branch that predicts the bounding box coordinates through a distribution focal loss (DFL) formulation and a classification branch that predicts class probabilities for targets. The regression output is processed as shown in equation (14):

$$B_i = \mathrm{Scale}_i\left(\mathrm{Conv}_{1\times1}(F_i'', 4 \times \mathrm{reg}_{\mathrm{max}})\right) \cdot \sqrt{\eta \cdot \mathrm{Norm}(F_i'')} \cdot \Gamma(\kappa) \tag{14}$$

where Scale_i represents a scale-specific adjustment factor for different feature levels, $\mathrm{reg}_{\mathrm{max}}$ denotes the number of bins for distribution focal loss, η is the feature normalization coefficient, and $\Gamma(\kappa)$ implements an adaptive scaling function with parameter κ controlling detection confidence

The classification process applies specialized weight matrices to the encoded feature representations, mapping them to appropriate class probability distributions and confidence scores for effective object categorization. The DFL methodology partitions each bounding box coordinate into $\mathrm{reg}_{\mathrm{max}}$ discrete bins, facilitating more precise localization of specific targets. During inference, the discrete probability distribution is converted to continuous coordinate values as expressed in Equation (15):

doi:10.20944/preprints202504.1498.v1

$$b_{\text{continuous}} = \sum_{j=0}^{\text{reg}_{\max}-1} P\left(b_j\right) \cdot j + \delta \cdot \sum_{j=0}^{\text{reg}_{\max}-1} P\left(b_j\right) \cdot \log(j+\epsilon) \cdot e^{\tanh(\rho \cdot j)} \tag{15}$$

where $P\left(b_j\right)$ represents the probability of the coordinate value falling within bin j, δ is the distribution refinement parameter, ϵ denotes a small constant ensuring numerical stability, and ρ serves as the bin importance modulation factor that enables sub-bin precision beyond the discrete quantization level. Comprehensive detection output integrates both classification and localization branches, combining class probability scores with spatial position information to generate complete detection results. Predicted box distributions undergo decoding into actual spatial coordinates utilizing anchor references and stride scaling factors as formulated in Equation (16):

$$\text{bbox} = \text{dist2bbox}(\text{DFL}(B), \text{anchors}) \times \text{strides} \tag{16}$$

Applying the new DEM architecture approach is a step to solve the target detection challenge in a resource-constrained environment with advanced convolution techniques, as traditional approaches sometimes disregard the useful textural and edge details. It allows for fast inference time for real time deployment because its parameter sharing is efficient and strong detection capabilities are maintained while reducing model size. Due to its good performance in detecting small objects, which are essential for long-range search and rescue, early target identification, this architecture is favorable. DEM's stable performance under various lighting and camouflage conditions is ensured by group normalization, and field tests confirm it to be superior in detecting camouflaged targets and at detecting distant personnel. It performs computationally efficiently in timing and accuracy, in the context of search and rescue operations, to support timely and accurate information and to optimize resource use.

4. Experiments

4.1. Experimental details and Evaluation criteria

The experiments were conducted on a system running Windows 10, with Python 3.10.16 and PyTorch 2.3.0. The hardware setup featured an RTX 3080 GPU, an Intel i7-11700K CPU, and CUDA version 11.8. For training, an SGD optimizer was used with a learning rate of 0.01 and a batch size of 32.

In the object detection tasks there are common performance metrics such as precision, recall, average precision (AP) and mean average precision (mAP). Osmosis is used to evaluate accuracy by precision, i.e. the ratio of correctly identified objects from all detected objects. Sensitivity (recall) is defined as proportion of positive samples that are correctly identified and is only a fraction of positive samples.

Often these two metrics show inverse relationship between them. Higher recall means that the model is indeed finding most of the true positives, but the precision will decrease because of false positives. High precision means that the model is likely to make correct predictions; however, high precision also means that some objects may be missing; hence, precision falls.

The Equation (17) – (20) for calculating recall, precision, AP, and mAP are as follows:

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

$$Precision = \frac{TP}{TP + FP} \tag{18}$$

$$AP = \int_0^1 P(r)dr \tag{19}$$

doi:10.20944/preprints202504.1498.v1

$$mAP = \frac{1}{n+1} AP_i \tag{20}$$

4.2. Datasets

In order to carry out research on efficient target detection in resource-constrained environments, we present a comprehensive dataset of 4616 images carefully chosen to enable solely tank and soldier detection in intricate environments as seen in Figure.5 and Figure.6. Then having a dataset which consist of real-world footage from the Russia Ukraine war and high-quality military simulation image that are close real battlefield but with controlled variation.

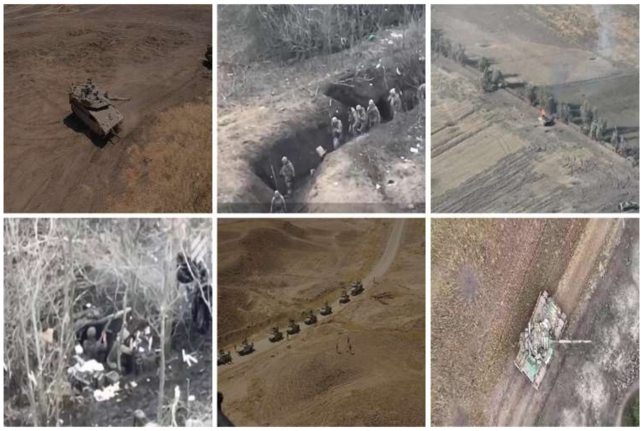
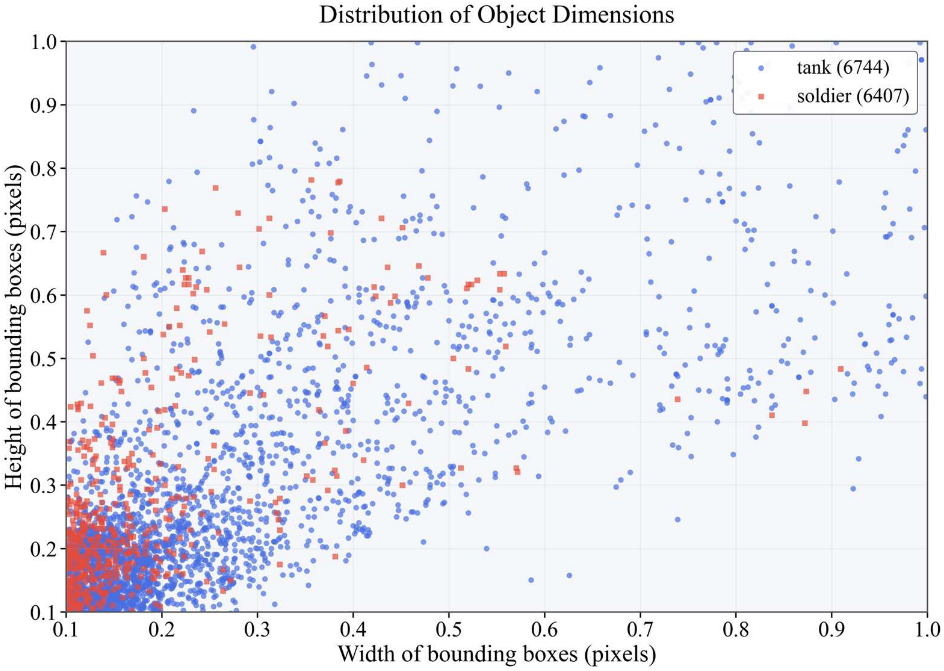


Figure 5. An example of dataset.

To create a challenging dataset, we have intentionally designed the dataset to include such detection scenarios as multi-scale targets, terrain occlusion, environmental obscurants (smoke and fog), advanced camouflage and image degradation (which is typical in reconnaissance video). The dataset concentrates on aerial views obtained from unmanned aerial vehicles (UAVs) and complemented by ground level perspectives, corresponding to the visual difficulties in performing modern complex environments.



doi:10.20944/preprints202504.1498.v1

Figure 6. The distribution path of each category of the dataset.

It is divided into 3 parts 3,231 images for training, 924 images for testing, and 461 images for validation. The purpose of this carefully assembled set is to serve as the basis for the creation of reliable detection algorithms to work in the harsh visual environment , resulting in better performance of the automated target detection systems.

4.3. Ablation study

This ablation study evaluates the contributions of three key modules in MSCDNet: the Multi-Scale Fusion Model (MSFM) as shown in Table 1, the Context Merge Module (CMM), and the Detail Enhance Module (DEM). The baseline model, with 2.58M parameters and 6.3G FLOPs, achieved an mAP50-95 of 38.2%. MSFM improved precision by 3.1% and mAP50-95 by 0.9%, while reducing parameters by 0.05M. CMM increased recall by 1.3% and mAP50-95 by 0.4%. DEM reduced parameters by 0.32M, decreased FLOPs by 0.3G, and increased mAP50-95 by 1.4%. The combinations of MSFM-CMM, MSFM-DEM, and CMM-DEM demonstrated further improvements: MSFM-CMM raised mAP50-95 by 1.3% and precision by 3.6%; MSFM-DEM achieved a 0.7% increase in mAP50-95 and reduced parameters by 0.32M; CMM-DEM reached 39.6% mAP50-95 with reduced computational demands.

Table 1. Ablation experimental results.

YOLO11n	MSFM	CMM	DEM	Param (M)	FLOPs (G)	P (%)	R (%)	mAP50 (%)	mAP50-95 (%)
√				2.58	6.3	81.8	66.5	71.5	38.2
√	√			2.53	6.3	84.9	65.8	72.5	39.1
√		√		2.59	6.3	82.4	67.8	72.6	38.6
√			√	2.26	6.0	83.4	67.6	72.5	39.6
√	√	√		2.54	6.3	85.5	66.7	73.7	39.5
√	√		√	2.21	6.0	84.1	67.9	73.6	38.9
√		√	√	2.26	6.0	83.7	67.0	73.2	39.6
√	√	√	√	2.22	6.0	86.1	68.1	74.7	40.1

The complete architecture incorporating all three modules achieved superior performance with precision increasing to 86.1%, recall improving to 68.1%, and mAP50-95 reaching 40.1%, while utilizing only 2.22M parameters and 6.0G FLOPs. These results validate our design approach, showing MSF enhances feature representation, CMM improves cross-modal information use, and DEM optimizes efficiency without sacrificing accuracy.

The PR curve comparison in Figure. 7 demonstrates the improved detection performance of our enhanced model. The curve for the improved model consistently maintains higher precision values across varying recall levels, particularly in the mid-to-high recall range (0.5-0.8), where the improved model shows significant advantage over the baseline. This indicates better confidence in predictions and fewer false positives while maintaining high recall, reflecting the synergistic benefits of the three modules working together.

The CAM (Class Activation Map) visualization in Figure. 8 reveals the attention mechanism differences between the baseline and improved models. The improved model demonstrates more focused and precise activation regions that closely align with the actual target objects, particularly highlighting discriminative features rather than background elements. This enhanced attention localization explains precision improvements, as the model more accurately concentrates on relevant target features while effectively suppressing background interference, a direct result of the MSF's improved feature representation and CMM's enhanced cross-modal information integration.

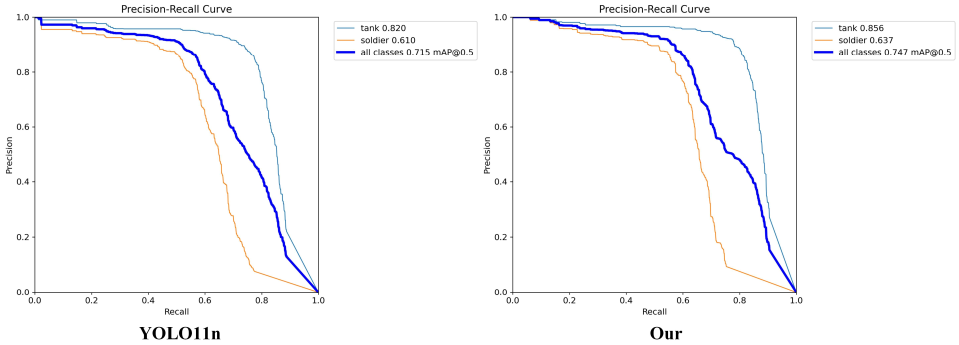


Figure 7. Comparison of PR curve before and after improvement.

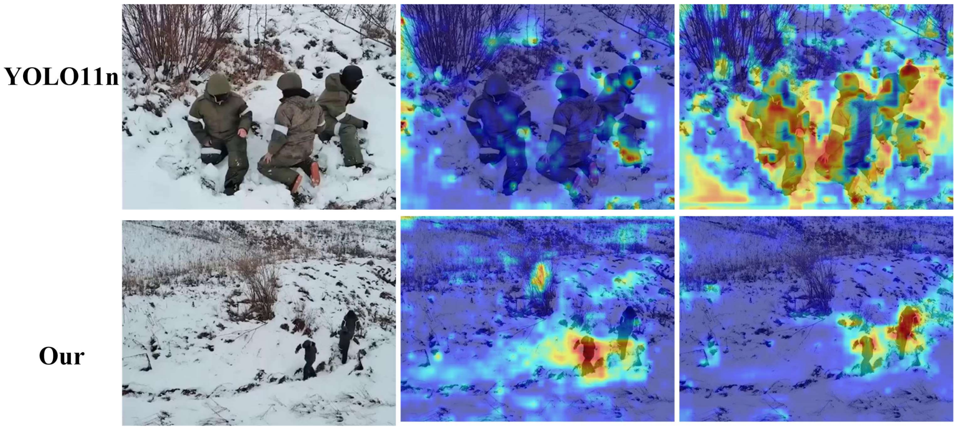


Figure 8. Comparison of the CAM effect of the model before and after the improvement.

4.4. Comparison with state-of-the-arts

MSCDNet demonstrates an exceptional efficiency-accuracy balance compared to contemporary object detection models as shown in Table 2 and Figure. 9. Traditional architecture like SSD show modest performance, with mAP50-95 8.3% lower than our model while requiring over 5 times more computational resources. More advanced models like DETR [50] and TOOD [51] offer improved accuracy but demand substantially higher computational resources, with TOOD requiring 33 times more FLOPs than our approach.

Recently lightweight models have provided more relevant comparisons. Our architecture outperforms RTMDet-Tiny [52] by 2.8% in mAP50-95 while using 25% fewer FLOPs and 54% fewer parameters. Similarly, it exceeds DFINE-n [53] and DEIM-n [54] variants by 2.6% in detection accuracy while reducing computational demands by 15% and parameter count by 40%.

The YOLO family shows progressive improvements across generations, yet our MSCDNet still demonstrates clear advantages. Compared to YOLOv5n [55], our model achieves 2.5% higher mAP50-95 with only slightly increased computational cost. Against YOLOv8n [56], we deliver a 2.3% accuracy improvement while reducing FLOPs by 26% and parameters by 26%. Compared to YOLOv10n [57], the closest competitor, our architecture improves mAP50-95 by 1.8% while reducing both parameter count and computational demands by nearly 27%. Our model also outperforms YOLOv11n by 1.9% in accuracy while using 4.8% fewer FLOPs and 14% fewer parameters.

Most notably, MSCDNet achieves balanced improvements across both precision and recall metrics, with values 4.3% and 1.6% higher than YOLOv11n respectively, demonstrating superior detection capability across diverse object classes and challenging scenarios.

doi:10.20944/preprints202504.1498.v1

Table 2. Compare the test results of different models.

Model	Precision	Recall	mAP50	mAP50-95	Flops	Param
	(%)	(%)	(%)	(%)	(G)	(M)
SSD	76.2	58.3	63.7	31.8	31.2	27.1
DETR	82.4	61.5	67.9	35.2	95.2	44.0
TOOD	83.5	63.7	70.2	37.1	199	32.04
RTMDet-Tiny	83.2	64.3	70.5	37.3	8.03	4.87
DFINE-n	82.8	64.5	70.6	37.5	7.12	3.73
DEIM-n	83.0	64.6	70.7	37.5	7.12	3.73
YOLOv5n	85.3	64.7	70.8	37.6	5.80	2.18
YOLOv8n	83.0	64.9	71.4	37.8	8.1	3.0
YOLOv10n	83.9	63.8	70.8	38.3	8.2	2.69
YOLOv11n	81.8	66.5	71.5	38.2	6.3	2.58
Our	86.1	68.1	74.7	40.1	6.0	2.22

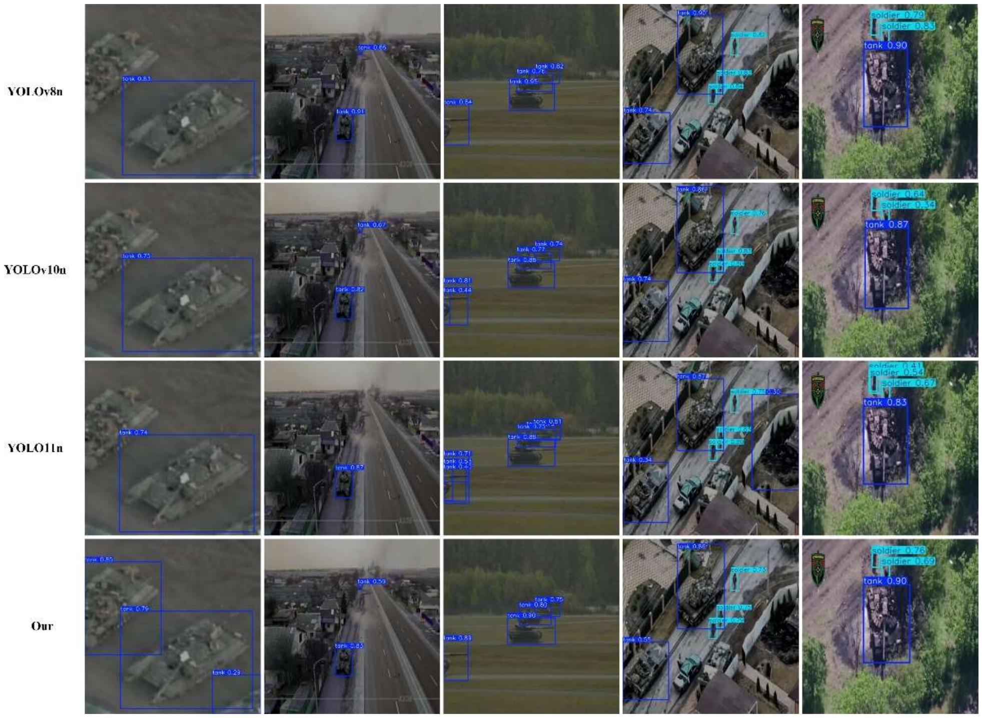


Figure 9. Detection effect of different models.

We evaluated various module configurations for battlefield object recognition within the YOLOv11 architecture, addressing challenges such as variable scales, complex morphologies, and cluttered backgrounds. Table 3 shows performance metrics for five YOLOv11 variants. The proposed MSFM module outperforms others with 84.9% precision, surpassing the C3K2 baseline by 3.1 percentage points, and maintains a competitive recall of 65.8%. It achieves the highest detection quality, with an mAP50 of 72.5% and an mAP50-95 of 39.1%. The C3k2-Star variant shows a slight precision improvement to 82.2%, but its recall drops significantly to 61.5%, limiting its effectiveness. The C3k2-IDWC variant strikes a balance with 83.0% precision and good computational efficiency at 6.1 GFLOPS, making it the most lightweight option. The MAN module achieves 83.2% precision but

doi:10.20944/preprints202504.1498.v1

struggles with recall at 62.2% and introduces higher computational costs at 8.4 GFLOPS and 3.77M parameters. MSFM stands out by providing superior detection performance without additional computational overhead, maintaining the same GFLOPS as the baseline C3K2, with only minor parameter differences, making it the optimal choice for object recognition in resource-constrained environments due to its balanced performance and efficiency.

Table 3. Comparative Analysis of YOLOv11 Module Variants for Object Recognition.

Model	Precision	Recall	mAP50	mAP50-95	Flops	Param
	(%)	(%)	(%)	(%)	(G)	(M)
C3K2	81.8	66.5	71.5	38.2	6.3	2.58
C3k2-Star [58]	82.2	61.5	68.3	35.5	6.4	2.47
C3k2-IDWC [59]	83.0	65.0	71.4	37.3	6.1	2.39
MAN [60]	83.2	62.2	70.2	37.1	8.4	3.77
MSFM	84.9	65.8	72.5	39.1	6.3	2.53

Table 4 presents the performance metrics of various FPN architectures, highlighting the advantages of our approach over existing methods. Traditional FPN shows significant limitations in military target detection. The Slimneck architecture exhibits suboptimal feature representation with a recall of 65.2%. BiFPN, while computationally efficient, struggles with complex targets, achieving a recall of 66.4%. MAFPN improves precision to 84.0%, but its recall drops to 65.1% and requires higher computational resources at 7.1 GFLOPS. Our architecture overcomes these limitations, achieving optimal precision of 82.4% and superior recall of 67.8%, surpassing Slimneck by 2.6%, BiFPN by 1.4%, and MAFPN by 2.7%. This improvement translates into better mean Average Precision metrics, with an mAP50 of 72.6% and mAP50-95 of 38.6%. Remarkably, our design maintains a moderate computational cost of 6.3 GFLOPS—comparable to BiFPN and 11.3% lower than MAFPN. This balance is achieved through the MSFM module, which enhances multi-scale feature extraction, and the CMM module, which enables advanced context-aware feature fusion. Together, these modules offer an efficient and high-performance solution for object recognition in resource-constrained environments.

Table 4. Performance Comparison of Various FPN Architectures in Object Recognition.

Model	Precision	Recall	mAP50	mAP50-95	Flops	Param
	(%)	(%)	(%)	(%)	(G)	(M)
Slimneck[61]	81.8	65.2	70.1	37.0	5.9	2.57
BiFPN[62]	82.0	66.4	70.9	37.6	6.3	1.92
MAFPN[63]	84.0	65.1	70.6	37.5	7.1	2.69
CMM	82.4	67.8	72.6	38.6	6.3	2.59

4.5. Generalization experiments

To assess the adaptability of our model across different domains, we evaluated its performance on two distinct datasets.

The VisDrone2019[64] dataset serves as a comprehensive benchmark for visual object detection in aerial imagery. It includes around 10,000 images and 288 video clips, with 2.6 million annotated objects spanning 10 categories. The dataset is divided into 6,471 images for training, 548 for validation, and 3,190 for testing.

The BDD100K[65] dataset presents challenges from a vehicle-mounted perspective. With over 100,000 images and 10 million annotations, it covers various weather conditions as well as day and night environments. For this dataset, we used 70,000 images for training, 10,000 for validation, and 20,000 for testing.

Table 5. Generalize experimental results on different datasets.

Dataset	Model	Precision (%)	Recall (%)	mAP50 (%)	mAP50-95 (%)
VisDrone	YOLO11n	45.5	33.4	33.7	19.6
	Our	45.7	35.1	34.8	20.1
BDD100K	YOLO11n	58.8	41.5	42.5	27.8
	Our	61.1	40.0	43.6	29.0

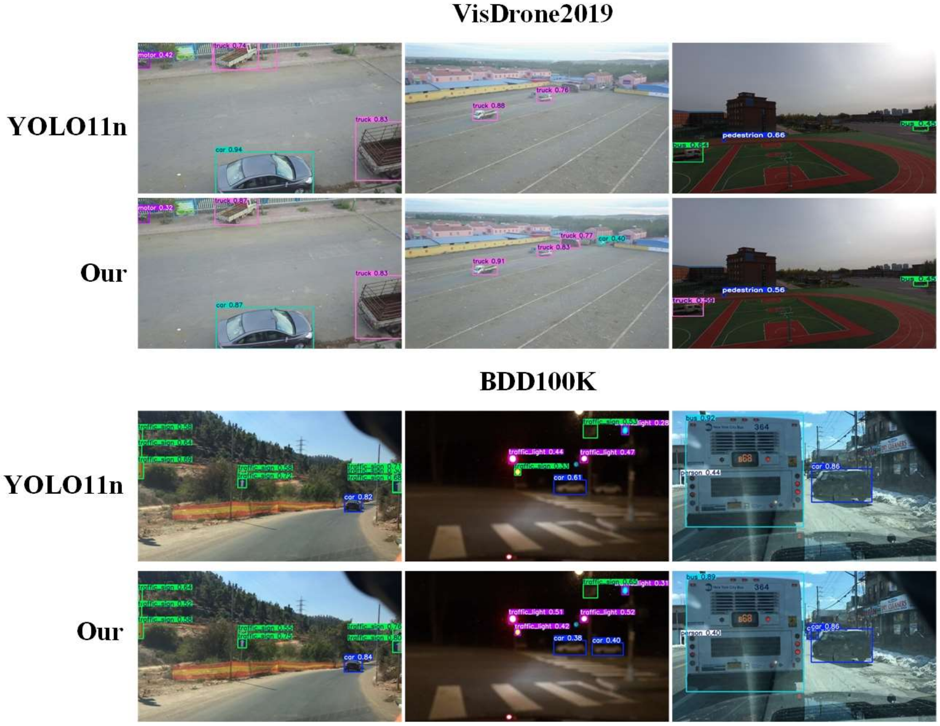


Figure 10. The effect of the model on VisDrone2019 and BDD100K datasets before and after is improved.

Comparing our model to the baseline YOLO11n, we observed consistent improvements across both datasets as shown in Table 5. On VisDrone, precision increased from 45.5% to 45.7%, a gain of 0.2%. Recall improved from 33.4% to 35.1%, an increase of 1.7%. For detection accuracy, mAP50 rose from 33.7% to 34.8%, a gain of 1.1%, and mAP50-95 improved from 19.6% to 20.1%, an increase of 0.5%. On BDD100K, precision improved from 58.8% to 61.1%, a gain of 2.3%. However, recall decreased from 41.5% to 40.0%, a reduction of 1.5%. Overall detection performance still improved, with mAP50 increasing from 42.5% to 43.6%, a gain of 1.1%, and mAP50-95 rising from 27.8% to 29.0%, an improvement of 1.2%.

These results highlight the superior detection performance of our model across different domains. Figure 10 demonstrates the improved detection effectiveness of our model before and after the enhancements, evaluated on both the VisDrone and BDD100K datasets, visually reinforcing our quantitative findings.

5. Conclusions

This paper introduces MSCDNet, a lightweight architecture for target detection in resource-constrained environments. Our approach integrates three key modules: Multi-Scale Fusion for enhanced feature representation, Context Merge Module for adaptive cross-scale integration, and

Detail Enhance Module for preserving critical details. Experiments demonstrate MSCDNet's superior performance with 40.1% mAP50-95, 86.1% precision, and 68.1% recall while requiring minimal computational resources of just 2.22M parameters and 6.0G FLOPs. Our model consistently outperforms contemporary architectures including YOLO variants while using fewer resources. Generalization tests across VisDrone2019 and BDD100K datasets confirm its effectiveness in diverse scenarios.

Despite these achievements, limitations remain in extreme weather conditions and Severe occlusion. Future work should explore cross-modal fusion for all-weather capability, adaptive computation mechanisms, self-supervised learning approaches, and hardware-aware optimizations to further enhance MSCDNet's applicability in resource-constrained environments where detection reliability directly impacts the application value in civilian fields such as personnel search and rescue, traffic management, etc.

Author Contributions: Conceptualization, K.W. and G.H.; methodology, K.W.; software, K.W.; validation, G.H., X.L.; formal analysis, K.W., X.L.; investigation, K.W., X.L.; resources, G.H.; data curation, G.H.; writing—original draft preparation, K.W.; writing—review and editing, K.W., G.H., X.L.; visualization, K.W.; supervision, G.H.; project administration, G.H.; funding acquisition, G.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Jia, L., & Pang, W. (2024). Overview of Battlefield Debris Data Fusion Technology for Situation Awareness. In 2024 Asia-Pacific Conference on Software Engineering, Social Network Analysis and Intelligent Computing (SSAIC) (pp. 474-478). IEEE.
2. Dehghan, M., & Khosravian, E. (2024). A Review of Cognitive UAVs: AI-Driven Situation Awareness for Enhanced Operations.
3. Tiwari, K. C., Arora, M. K., & Singh, D. (2011). An assessment of independent component analysis for detection of military targets from hyperspectral images. International Journal of Applied Earth Observation and Geoinformation, 13(5), 730-740.
4. Palm, H. C., Ajer, H., & Haavardsholm, T. V. (2008). Detection of military objects in LADAR images.
5. Deveci, M., Kuvvetli, Y., & Akyurt, İ. Z. (2020). Survey on military operations of fuzzy set theory and its applications. Journal of Naval Sciences and Engineering, 16(2), 117-141.
6. Riedl, J. L. (1976). CCD Sensor Array and Microprocessor Application to Military Missile Tracking. In Modern Utilization of Infrared Technology II (Vol. 95, pp. 148-154). SPIE.
7. Schaber, G. G. (1999). SAR studies in the Yuma Desert, Arizona: sand penetration, geology, and the detection of military ordnance debris. Remote sensing of environment, 67(3), 320-347.
8. Lv, J., Zhu, D., Geng, Z., Han, S., Wang, Y., Yang, W., ... & Zhou, T. (2023). Recognition of deformation military targets in the complex scenes via MiniSAR submeter images with FASAR-Net. IEEE Transactions on Geoscience and Remote Sensing, 61, 1-19.
9. Tiwari, K. C., Arora, M. K., Singh, D., & Yadav, D. (2013). Military target detection using spectrally modeled algorithms and independent component analysis. Optical Engineering, 52(2), 026402-026402.
10. Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. arXiv preprint arXiv:1602.07360.

11. Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).
12. Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence, 39(6), 1137-1149.
13. He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).
14. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14 (pp. 21-37). Springer International Publishing.
15. Hussain, M. (2023). YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection. Machines, 11(7), 677.
16. Sun, Y., Liu, Z., Todorovic, S., & Li, J. (2007). Adaptive boosting for SAR automatic target recognition. IEEE Transactions on Aerospace and Electronic Systems, 43(1), 112-125.
17. Zhang, L., Shi, Z., & Wu, J. (2015). A hierarchical oil tank detector with deep surrounding features for high-resolution optical satellite imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8(10), 4895-4909.
18. Zhang, L., Chu, R., Xiang, S., Liao, S., & Li, S. Z. (2007). Face detection based on multi-block lbp representation. In Advances in Biometrics: International Conference, ICB 2007, Seoul, Korea, August 27-29, 2007. Proceedings (pp. 11-18). Springer Berlin Heidelberg.
19. Pei, J., Huang, Y., Sun, Z., Zhang, Y., Yang, J., & Yeo, T. S. (2018). Multiview synthetic aperture radar automatic target recognition optimization: Modeling and implementation. IEEE Transactions on Geoscience and Remote Sensing, 56(11), 6425-6439.
20. Che, J., Fang, L., Zhong, Z., Su, X., Ma, Q., & Yu, G. (2024, July). A survey of automatic target recognition technology based on multi-source data fusion. In IET Conference Proceedings CP886 (Vol. 2024, No. 12, pp. 1592-1599). Stevenage, UK: The Institution of Engineering and Technology.
21. Li, Y., Luo, Y., Zheng, Y., Liu, G., & Gong, J. (2024). Research on Target Image Classification in Low-Light Night Vision. Entropy, 26(10), 882.
22. Salmon, P. M., Lenné, M. G., Triggs, T., Goode, N., Cornelissen, M., & Demczuk, V. (2011). The effects of motion on in-vehicle touch screen system operation: A battle management system case study. Transportation research part F: traffic psychology and behaviour, 14(6), 494-503.
23. Bajracharya, M., Moghaddam, B., Howard, A., Brennan, S., & Matthies, L. H. (2009). A fast stereo-based system for detecting and tracking pedestrians from a moving vehicle. The International Journal of Robotics Research, 28(11-12), 1466-1485.
24. Chaves, S. M. (2010). Using Kalman filtering to improve a low-cost GPS-based collision warning system for vehicle convoys.
25. Boult, T. E., Micheals, R. J., Gao, X., & Eckmann, M. (2001). Into the woods: Visual surveillance of noncooperative and camouflaged targets in complex outdoor settings. Proceedings of the IEEE, 89(10), 1382-1402.
26. Zhao, Q. (2021). Aboveground Storage Tank Detection Using Faster R-CNN and High-Resolution Aerial Imagery (Master's thesis, Duke University).
27. Peng, S. (2024). Multi-object extraction technology for complex background based on faster regions-CNN algorithm in the context of artificial intelligence. Service Oriented Computing and Applications, 1-13.
28. Naz, P., Hengy, S., & Hamery, P. (2012, May). Soldier detection using unattended acoustic and seismic sensors. In Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR III (Vol. 8389, pp. 183-194). SPIE.
29. Xu, D., & Wu, Y. (2020). Improved YOLO-V3 with DenseNet for multi-scale remote sensing target detection. Sensors, 20(15), 4276.
30. Wang, H., Qian, H., Feng, S., & Wang, W. (2024). L-SSD: lightweight SSD target detection based on depth-separable convolution. Journal of Real-Time Image Processing, 21(2), 33.
31. Wang, S., Du, Y., Zhao, S., & Gan, L. (2023). Multi-scale infrared military target detection based on 3X-FPN feature fusion network. IEEE Access, 11, 141585-141597.

32. Liu, J., Jia, R., Li, W., Ma, F., Abdullah, H. M., Ma, H., & Mohamed, M. A. (2020). High precision detection algorithm based on improved RetinaNet for defect recognition of transmission lines. Energy Reports, 6, 2430-2440.
33. Pushkarenko, Y., & Zaslavskyi, V. (2024). Research on the state of areas in Ukraine affected by military actions based on remote sensing data and deep learning architectures. Radioelectronic and Computer Systems, 2024(2), 5-18.
34. Li, T., Wang, H., Li, G., Liu, S., & Tang, L. (2022, June). SwinF: Swin Transformer with feature fusion in target detection. In Journal of Physics: Conference Series (Vol. 2284, No. 1, p. 012027). IOP Publishing.
35. Zhuang, X., Li, D., Wang, Y., & Li, K. (2024). Military target detection method based on EfficientDet and Generative Adversarial Network. Engineering Applications of Artificial Intelligence, 132, 107896.
36. Sun, Y., Wang, J., You, Y., Yu, Z., Bian, S., Wang, E., & Wu, W. (2025). YOLO-E: a lightweight object detection algorithm for military targets. Signal, Image and Video Processing, 19(3), 241.
37. Jani, M., Fayyad, J., Al-Younes, Y., & Najjaran, H. (2023). Model compression methods for YOLOv5: A review. arXiv preprint arXiv:2307.11904.
38. Zhang, W., Jiao, L., Liu, X., & Liu, J. (2019, July). Multi-scale feature fusion network for object detection in VHR optical remote sensing images. In IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium (pp. 330-333). IEEE.
39. Fan, L., Wang, H., Yang, Q., Chen, X., Deng, B., & Zeng, Y. (2022). Fast detection and reconstruction of tank barrels based on component prior and deep neural network in the terahertz regime. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-17.
40. Ma, C., Zhang, Y., Guo, J., Hu, Y., Geng, X., Li, F., ... & Ding, C. (2021). End-to-end method with transformer for 3-D detection of oil tank from single SAR image. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-19.
41. Song, Z., Kang, X., Wei, X., Dian, R., Liu, J., & Li, S. (2024). Multi-granularity Context Perception Network for Open Set Recognition of Camouflaged Objects. IEEE Transactions on Multimedia.
42. Naeem, W., Sutton, R., & Xu, T. (2012). An integrated multi-sensor data fusion algorithm and autopilot implementation in an uninhabited surface craft. Ocean Engineering, 39, 43-52.
43. Lv, J., Zhu, D., Geng, Z., Han, S., Wang, Y., Yang, W., ... & Zhou, T. (2023). Recognition of deformation military targets in the complex scenes via MiniSAR submeter images with FASAR-Net. IEEE Transactions on Geoscience and Remote Sensing, 61, 1-19.
44. Wu, B., & Zhou, J. (2024). Video-Based Martial Arts Combat Action Recognition and Position Detection Using Deep Learning. IEEE Access.
45. Choudhary, S. (2023). Real time pixelated camouflage texture generation (Doctoral dissertation, School of Computer Science, UPES, Dehradun).
46. Barnawi, A., Budhiraja, I., Kumar, K., Kumar, N., Alzahrani, B., Almansour, A., & Noor, A. (2022). A comprehensive review on landmine detection using deep learning techniques in 5G environment: Open issues and challenges. Neural Computing and Applications, 34(24), 21657-21676.
47. Anzer, G., Bauer, P., Brefeld, U., & Faßmeyer, D. (2022). Detection of tactical patterns using semi-supervised graph neural networks. In 16th MIT sloan sports analytics conference (pp. 1-15).
48. Wang, Q., Fu, M., Wang, J., Sun, L., Huang, R., Li, X., ... & Jiang, C. (2024). Free-walking: Pedestrian inertial navigation based on dual foot-mounted IMU. Defence Technology, 33, 573-587.
49. Jocher, G. YOLO11. 2024. Available online: https://github.com/ultralytics/ultralytics/tree/main.
50. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Computer Vision—ECCV 2020; Springer International Publishing: Cham, Switzerland, 2020; pp. 213–229.
51. Feng, C.; Zhong, Y.; Gao, Y.; Scott, M.R.; Huang, W. Tood: Task-aligned one-stage object detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal QC, Canada, 10–17 October 2021; pp. 3490–3499.
52. Lyu, C.; Zhang, W.; Huang, H.; Zhou, Y.; Wang, Y.; Liu, Y.; Zhang, S.; Chen, K. Rtmdet: An empirical study of designing real-time object detectors. arXiv 2022, arXiv:2212.07784.

53. Peng, Y., Li, H., Wu, P., Zhang, Y., Sun, X., & Wu, F. (2024). D-FINE: redefine regression Task in DETRs as Fine-grained distribution refinement. arXiv preprint arXiv:2410.13842.
54. Huang, S., Lu, Z., Cun, X., Yu, Y., Zhou, X., & Shen, X. (2024). DEIM: DETR with Improved Matching for Fast Convergence. arXiv preprint arXiv:2412.04234.
55. Jocher, G., Stoken, A., Borovec, J., Changyu, L., Hogan, A., Diaconu, L., ... & Dave, P. (2020). ultralytics/yolov5: v3. 0. Zenodo.
56. Glenn, J. YOLOv8. 2023. Available online: https://github.com/ultralytics/ultralytics/tree/main.
57. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. Yolov10: Real-time end-to-end object detection. arXiv 2024, arXiv:2405.14458.
58. Ma, X., Dai, X., Bai, Y., Wang, Y., & Fu, Y. (2024). Rewrite the stars. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5694-5703).
59. Yu, W., Zhou, P., Yan, S., & Wang, X. (2024). Inceptionnext: When inception meets convnext. In Proceedings of the IEEE/cvf conference on computer vision and pattern recognition (pp. 5672-5683).
60. Feng, Y., Huang, J., Du, S., Ying, S., Yong, J. H., Li, Y., ... & Gao, Y. (2024). Hyper-yolo: When visual object detection meets hypergraph computation. IEEE Transactions on Pattern Analysis and Machine Intelligence.
61. Li, H., Li, J., Wei, H., Liu, Z., Zhan, Z., & Ren, Q. (2022). Slim-neck by GSConv: A better design paradigm of detector architectures for autonomous vehicles. arXiv preprint arXiv:2206.02424, 10.
62. Chen, J., Mai, H., Luo, L., Chen, X., & Wu, K. (2021, September). Effective feature fusion network in BIFPN for small object detection. In 2021 IEEE international conference on image processing (ICIP) (pp. 699-703). IEEE.
63. Yang, Z., Guan, Q., Zhao, K., Yang, J., Xu, X., Long, H., & Tang, Y. (2024, October). Multi-branch Auxiliary Fusion YOLO with Re-parameterization Heterogeneous Convolutional for Accurate Object Detection. In Chinese Conference on Pattern Recognition and Computer Vision (PRCV) (pp. 492-505). Singapore: Springer Nature Singapore.
64. Du, D., Zhu, P., Wen, L., Bian, X., Lin, H., Hu, Q., ... & Zhang, L. (2019). VisDrone-DET2019: The vision meets drone object detection in image challenge results. In Proceedings of the IEEE/CVF international conference on computer vision workshops (pp. 0-0).
65. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., ... & Darrell, T. (2020). Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2636-2645).

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.

Statement: Current research is limited to the Lightweight Computer Vision Algorithms in Object Detection, the main value of this research is to promote the application ability of computer vision technology in resource-constrained environments, which is of great value to many civil fields such as disaster relief, traffic safety, and does not pose a threat to public health or national security. Authors acknowledge the dual use potential of the research involving the target detection technology and confirm that all necessary precautions have been taken to prevent potential misuse. As an ethical responsibility, authors strictly adhere to relevant national and international laws about DURC. Authors advocate for responsible deployment, ethical considerations, regulatory compliance, and transparent reporting to mitigate misuse risks and foster beneficial outcomes.