This ACCV 2024 paper, provided here by the Computer Vision Foundation, is the author-created version.
The content of this paper is identical to the content of the officially published ACCV 2024
LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/accv

ESM-YOLO: Enhanced Small Target Detection Based on Visible and Infrared Multi-modal Fusion

Qianqian Zhang[1,2], Linwei Qiu[3], Li Zhou[1]*, and Junshe An[1,4]

[1] National Space Science Center, Chinese Academy of Sciences, Beijing 101499, China
[2] School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China
[3] School of Astronautics, Beihang University, Beijing 100191, China
[4] School of Astronomy and Space Science, University of Chinese Academy of Sciences, Beijing 100049, China
zhangqianqian21@mails.ucas.ac.cn, qiulinwei@buaa.edu.cn,
{zhouli,anjunshe}@nssc.ac.cn

Abstract. Detecting small targets in remote sensing imagery is frequently impeded by target faintness and complex background, resulting in reduced accuracy. This work introduces an Enhanced Small Target Detection method, termed **ESM-YOLO**, which leverages multi-modal fusion of visible and infrared data to enhance inter-modality correlation and thereby augments performance. Firstly, we devise a pixel-level Bilateral Excitation Fusion (BEF) module to extract both shared and unique features from distinct modalities symmetrically and efficiently. Subsequently, an Improved Atrous Spatial Pyramid Pooling (IASPP) unit and a Compact BottleneckCSP (CBCSP) unit are incorporated into the detection architecture. These components are meticulously tailored to enhance the detection of minute object features, while ensuring a balance between computational efficiency and feature representation capability. Experimental results show that ESM-YOLO achieves 82.42% accuracy on the widely used Vehicle Detection in Aerial Imagery (VEDAI) dataset. The effectiveness and superiority of our proposed method are demonstrated through extensive experiments.

Keywords: Multi-modal fusion · Visible light images · Infrared images · Small Targets

1 Introduction

The advancements in artificial intelligence technology have significantly impacted target detection techniques utilized in satellites and unmanned aerial vehicles (UAVs), with profound implications across both civilian and military sectors. Civilian applications encompass diverse areas such as civil vehicle tracking [31] and city target detection [29]. In the military realm, a pivotal application is the detection of military camouflage targets [22]. However, remote sensing images pose unique challenges due to their small and densely packed targets, intricate backgrounds, and substantial interference, resulting in compromised

target detection and recognition accuracy, along with unsatisfactory false detection and false alarm rates. The technical challenges associated with small remote sensing target detection render traditional methods insufficient. This limitation stems primarily from inadequate feature information extraction, fusion capabilities, and limited feature expression within the network. Consequently, enhancing the accuracy of remote sensing small target recognition has garnered significant attention as a pivotal research area.

Visible light imaging systems, which rely on external illumination, inherently capture rich color and texture information about targets, making them adept at imaging under favorable lighting conditions. However, in scenarios involving localized strong light, backlight, or extreme weather phenomena such as rain, snow, and smog, recognition accuracy suffers. Conversely, infrared imaging systems, leveraging thermal radiation characteristics for passive detection, exhibit resilience against interference in night and low-light environments, offering robust detection capabilities. Nevertheless, infrared images are inherently limited in shape and texture information, posing challenges for standalone small target recognition.

Advancements in imaging technology have facilitated the utilization of multispectral cameras to acquire remote sensing images, enabling the exploitation of complementary information across multiple modalities. This approach has the potential to enhance detection accuracy by fusing data from different spectral bands. Zhang *et al.* [28] investigated small target recognition and achieved improved results through the fusion of visible and infrared modalities, demonstrating the effectiveness of multimodal data integration in this context.

The main difficulties in multi-modal fusion detection are including scale diversity, differential feature fusion, alignment and misalignment of different modal data, loss of feature information, noise interference, computational resource limitations, and scarcity of training data. The cause of scale diversity is that small targets may exhibit different scale properties in different modal data. Differential feature fusion means that there may be differences in the representation of features such as the color and texture of small targets in different modal data, and effective fusion and extraction of useful information is the key. During the acquisition of multi-modal data, the spatial alignment and rectification of small targets may be inaccurate due to differences in acquisition viewpoints and time leading to suboptimal data fusion effect. The main cause of feature information loss and noise interference is that small targets occupy few pixels and are easily lost during feature extraction. Multimodal algorithms require to realize data fusion from different sensors, which can increase the complexity and computational cost of the algorithm. Small target detection models in remote sensing images are also limited by computational resources since they are mainly deployed on satellites and UAVs for real-time processing. In addition, the labeling of small target data in multi-modal scenarios requires significant financial and time costs, and currently available multi-modal small target datasets are relatively scarce. Therefore, it is an important work to study small target detection based on multi-modal fusion.

Multi-modal fusion algorithms are classified into three primary types based on the fusion level: pixel-level fusion [13], feature-level fusion [26], and decision-level fusion [25]. Pixel-level fusion involves the integration of raw data from various sensors during the preprocessing stage, ensuring minimal alteration of the original information. Feature-level fusion, in contrast, merges high-level abstracted features, which can potentially result in feature loss and consequently missed detections. Decision-level fusion, while effective, may encounter the issue of double-counting across different modal branches, leading to increased computational demands. Given these considerations, pixel-level fusion is deemed more appropriate for satellite or airborne applications due to its proximity to the data source and its ability to maintain data integrity throughout the fusion process.

Due to the diverse characteristics of multi-modal data, it is difficult to achieve ideal results by directly applying generalized unimodal detection algorithms to infrared and visible multi-modal fusion small target detection scenarios. Therefore, we propose ESM-YOLO, an enhanced small target detection method for visible and infrared multi-modal fusion. The pixel-level fusion algorithm of Bilateral Excitation Fusion (BEF) module is introduced to integrate information from both RGB and infrared (IR) images. This algorithm operates at the granularity of individual pixels, ensuring a meticulous combination of color-rich visual details from RGB data with the thermal and penetrative properties extracted from IR imagery. Additionally, addressing the challenges of scale variability and the need for effective feature fusion in multi-modal small target detection, we employ the Improved Atrous Spatial Pyramid Pooling (IASPP) structure to harmoniously merge diverse feature information. Owing to the minute presence of targets in remote sensing images and the consequent risk of feature information loss, our strategy bolsters the network's feature extraction potency through the integration of the Compact BottleneckCSP (CBCSP) structure within the Head module. Comprehensive experimentation conducted on the VEDAI remote sensing image dataset has validated the efficacy and superiority of our proposed enhancement in comparison to prevailing methodologies. The main work of this paper is as follows:

- The Bilateral Excitation Fusion (BEF) module is incorporated to fortify the network's multi-modal integration capabilities and enhance its cross-modal learning capacity.
- The Improved Atrous Spatial Pyramid Pooling (IASPP) structure is devised to augment the feature extraction efficacy of the network. Furthermore, a Compact BottleneckCSP (CBCSP) structure, integrated within the Head component, is proposed specifically for small target detection, thereby leading to a substantial boost in the model's overall detection capabilities.
- Extensive experiments based on the VEDAI remote sensing image dataset have demonstrated the effectiveness and superiority of the proposed ESM-YOLO method over existing methods.

2 Related Work

2.1 Object Detection Using Multimodal Data

Multimodal fusion, a technique that integrates various types of sensing data, is instrumental in detecting objects or targets within real-world settings. This approach has gained widespread application in diverse scenarios, such as autonomous vehicle navigation [5] and medical retina analysis [1]. Studies have demonstrated that multimodal fusion can concurrently identify multiple targets with high precision and a minimal false alarm rate, outperforming single-modal methods. In the realm of small target detection, the amalgamation of data from multiple sensors—encompassing RGB imagery, synthetic aperture radar (SAR), light detection and ranging (LiDAR), infrared (IR), panchromatic (PAN), and multispectral (MS) data—significantly bolsters the efficacy of detection operations. Kang *et al.* [10] introduced a real-time framework for remote sensing target detection that leverages visible and infrared multimodal data. Their experiments substantiated the model's capability to accurately discern vehicle targets across varying illumination conditions, markedly reducing both detection omissions and false positives when compared to single-modal algorithms. Despite the performance enhancements afforded by multimodal fusion algorithms, challenges persist, such as the need for improved detection accuracy and the substantial computational demands required to meet the stringent timeliness criteria of real-time detection tasks.

2.2 Small Target Detection

Small target detection is a vital yet challenging task in the realm of computer vision and image processing, particularly significant in applications such as surveillance [2,12,15], remote sensing [27,30], and autonomous navigation systems [7]. The acquisition of remote sensing data is marked by its multi-scale, multi-angle, and multi-target nature, posing greater complexities for multi-modal small target detection compared to conventional target detection tasks. To address these challenges, researchers have devised various advancements, encompassing the employment of data enhancement techniques, the refinement of feature extraction networks, the enhancement of feature fusion networks, and the exploration of diverse combinations of network architectures and anchor frame algorithms. These strategies are designed to enhance the precision of network-based target detection and minimize the false alarm rate.

The backbone feature extraction network plays a pivotal role in discerning small targets amidst intricate backgrounds. The network's feature extraction prowess is a direct determinant of the recognition system's precision. Zhang *et al.* [30] introduced a novel approach by integrating a Transformer [21] backbone within the BottleNeckCSP [24] framework. This innovation aims to augment the capture of global information and refine the feature extraction for small objects, thereby mitigating the challenges associated with the detection of poorly visible small targets. The function of a feature fusion network is to amalgamate data

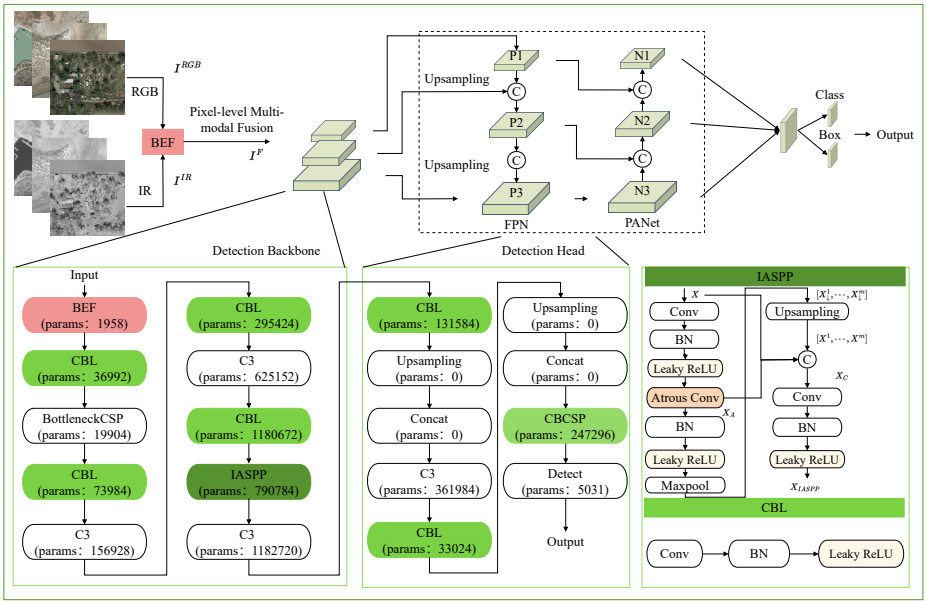

Fig. 1: Overall framework of the ESM-YOLO model. It consists of three modules: pixel-level multi-modal fusion, *i.e.* Bilateral Excitation Fusion Module (BEF), detection backbone, and detection head.

from various sensors and sources, thereby enhancing the precision and reliability of target detection. In their research, Yu *et al.* [27] introduced an enhancement to the spatial pyramid pooling (SPP) module by substituting the conventional maximum pooling with atrous convolution.

3 Methods and Datasets

This section is divided into four parts, describing the proposed structures, loss functions, the visible infrared multi-modal dataset VEDAI [17], and the evaluation metrics used to validate the methodology in this paper.

3.1 Overall Framework of the ESM-YOLO Model

This section outlines the architecture and specific design elements of our enhanced methodology. Informed by prior analyses, we introduce a Multi-Modal Fusion Real-time Small Target Recognition Network tailored for remote sensing scenarios, depicted in Fig. 1. Dubbed ESM-YOLO, the network is structured around three core components: a multi-modal fusion module, a backbone, and a head inclusive of a neck. Initiating the process, the Bilateral Excitation Fusion Module (BEF) integrates visible and infrared image modalities at the pixel level, yielding a fused input. The backbone is tasked with extracting both

low-level textural and high-level semantic features from this fused input. These features are subsequently directed to the Head network. Within the Head, feature fusion occurs via a feature pyramid network (FPN) framework, generating multi-resolution feature maps. Anchored boxes, predetermined in size, guide the formation of bounding boxes for target prediction. Ultimately, the network concludes with classification and localization operations to effectuate target detection and output recognition results, thereby completing the real-time small target identification pipeline in remote sensing contexts.

Bilateral Excitation Fusion Module The Bilateral Excitation Fusion (BEF) Module is designed to integrate information from both visible light (RGB) images and infrared (IR) images. This fusion of modalities at pixel level aims to capitalize on the complementary strengths of each type of image, thereby enhancing the overall feature representation and improving performance. Our BEF can modulate the features based on both spatial similarity and feature similarity. Spatially, it considers neighboring pixels to preserve local structures. In terms of feature similarity, it adjusts the contributions of different channels (*e.g.*, R, G, B, and IR intensity) based on their relevance, which enhances the discriminative power of the fused features for better detection. The whole process of BEF is described as follows.

Given the input RGB image $I^{RGB} \in \mathbb{R}^{C\times H\times W}$ and infrared modality $I^{IR} \in \mathbb{R}^{1\times H\times W}$, the scaled spatial information between RGB modality and IR modality are obtained by

$$I_1^{RGB} = I^{RGB} \cdot p^{RGB}, \quad I_1^{IR} = I^{IR} \cdot p^{IR}, \tag{1}$$

where p^{RGB} and p^{IR} are two weight parameters. Then, the internal spatial information $I_2^{RGB} \in \mathbb{R}^{C_1\times h\times w}$, $I_2^{IR} \in \mathbb{R}^{C_2\times h\times w}$ among two modalities are acquired by

$$I_2^{RGB} = \mathrm{Conv}_{1\times 1}(I_1^{RGB} \otimes \mathrm{Conv}_{1\times 1}(I_1^{RGB}) + I^{RGB}), \tag{2}$$

$$I_2^{IR} = \mathrm{Conv}_{1\times 1}(I_1^{IR} \otimes \mathrm{Conv}_{1\times 1}(I_1^{IR}) + I^{IR}), \tag{3}$$

where $\mathrm{Conv}_{1\times 1}$ represents the 1×1 convolution, and $\otimes$ is the element-wise multiplication. Along the spatial dimension, feature compression is executed subsequently, converting each two-dimensional feature channel from multiple modalities into a modal factor $M \in \mathbb{R}^{(C_1+C_2)\times 1\times 1}$, *i.e.*

$$M = \sigma_S(\mathrm{FC}(\sigma_R(\mathrm{FC}(M')))), M' = \mathrm{GAP}(\mathrm{Cat}[I_2^{RGB}, I_2^{IR}]) \tag{4}$$

where FC is the fully connected layer, GAP is the global average pooling, and σ_R, σ_S are ReLU, Sigmoid functions respectively. The factor M encapsulates a global receptive field to a certain degree, with its output dimension matching the input feature channels' count. It embodies the global response distribution across channels, empowering layers near the input to acquire a broader context-awareness by extending their receptive fields, ultimately enhancing their understanding of contextual information. Eventually, the final fusion results is

$$I^F = M \cdot (\mathrm{Cat}[I_2^{RGB}, I_2^{IR}]). \tag{5}$$

The integration of RGB and IR imagery via the BEF module capitalizes on their complementary properties, augmenting the system's perceptual capabilities, especially in intricate environments where small targets are prevalent.

Improved Atrous Spatial Pyramid Pooling To improve the model's ability to recognize objects of different sizes, we design the Improved Atrous Spatial Pyramid Pooling (IASPP) unit as seen in Fig. 1. In the novel architecture, it harnesses atrous convolutions to perform multi-scale feature extraction via an upgraded maximum pooling mechanism. Upsampling and channel concatenation strategies are employed to integrate these multi-scale features, thereby enhancing the network's capacity to discern objects across a range of sizes. To align features for seamless integration, upsampled pooled feature maps, achieved through nearest-neighbor interpolation, ensure congruent spatial dimensions across scales. To bolster feature expressiveness, Leaky ReLU activation (denoted as σ_L) follows each convolutional layer, introducing crucial nonlinearity. Recognizing the computational burden of these enhancements, grouped convolutions within the atrous layers, with the number of groups matched to the input channel count, are adopted to curtail computational overhead effectively.

The IASPP module is implemented as following. Given the feature map input X, we can obtain astrous features X_A by

$$X_A = \mathrm{AC}[\mathrm{CBL}(X)], \tag{6}$$

where AC denotes atrous convolution and CBL comprises `Conv-BN-Leaky ReLU` block. Subsequently, the module sequence denoted as m incorporates a max pooling operation that acts on the input X_A, thereby extracting features across various scales and consolidating them into pooled features, depicted as follows

$$X_{\downarrow}^{i} = \mathrm{MaxPool}_{\downarrow i}\left(\sigma_L\left(\mathrm{BN}\left(X_A\right)\right)\right), i = 1, \cdots, m. \tag{7}$$

Next, each pooled feature map $[X_{\downarrow}^{1}, \cdots, X_{\downarrow}^{m}]$ undergoes nearest-neighbor interpolation for upsampling, aligning its spatial dimensions with those of the initial input feature map, *i.e.* $[X^1, \cdots, X^m]$. Subsequently, the original input feature map X, the outcome of the atrous convolution X_A, and the channel-wise upsampled pooled feature maps $[X^1, \cdots, X^m]$ are concatenated along the channel axis. This operation generates a fused feature representation X_C, as detailed in Eq. (8).

$$X_C = \mathrm{Cat}\left[X, X_A, X^1, \cdots, X^m\right]. \tag{8}$$

Finally, the concatenated feature map X_C, is processed through another CBL to yield the ultimate output X_{IASPP}.

Compact BottleneckCSP Given an input feature X', the final results through Compact BottleneckCSP (CBCSP) module is

$$X_{CBCSP} = \mathrm{CBL}\left(\sigma_L\left(\mathrm{BN}\left(X_1'\right)\right)\right), X_1' = \mathrm{Cat}\left[\mathrm{Bottleneck}\left(\mathrm{CBL}(X')\right), \mathrm{CBL}(X')\right], \tag{9}$$

where Bottleneck is the Bottleneck structure in YOLOv5 [19]. The advantages of introducing the CBCSP for small target recognition and detection are mainly

reflected in the following aspects. First, CBCSP integrates features across multiple resolutions, preserving a broader feature spectrum. This is vital for small targets with weak feature signatures, as fusion of varied scale feature maps amplifies their representation, enhancing detection precision. Second, Incorporating residual links and dimensionality reduction, CBCSP's bottleneck design reduces computational load while boosting computational efficiency. Third, The CSP layer fosters adaptive and nonlinear representation capabilities, enhancing the model's versatility under diverse scenes and lighting, thereby increasing detection robustness. Lastly, Through multi-level feature map fusion, CBCSP equips the model to extract rich contextual and structural details about targets, facilitating more accurate identifications.

3.2 Loss Function

The detection loss $\mathcal{L}$ encompasses three integral components, *i.e.* object presence loss $\mathcal{L}_o$, object localization loss $\mathcal{L}_l$, and object classification loss $\mathcal{L}_c$ [19], collectively assessing the discrepancy between predictions and ground truth as expressed in Eq. (10).

$$\mathcal{L} = \lambda_o \sum_{a=0}^{2} \alpha_o^a \mathcal{L}_o + \lambda_l \sum_{a=0}^{2} \alpha_l^a \mathcal{L}_l + \lambda_c \sum_{a=0}^{2} \alpha_c^a \mathcal{L}_c, \tag{10}$$

where a denotes the index of the output layer within the head; α_o^a, α_l^a, and α_c^a denote the respective weights assigned to the object presence, localization, and classification losses for each layer; and the weights λ_o, λ_l, and λ_c serve as global weights, tuning the emphasis on errors pertaining to object presence, absence of objects, bounding box coordinates, dimensions, and class categorization.

3.3 Datasets

Our experiments utilize the widely recognized Vehicle Detection in Aerial Imagery (VEDAI) dataset [17], which is derived from a subset of the expansive Utah Automated Geographic Reference Center (AGRC) dataset. The AGRC dataset comprises images captured from a uniform altitude, each spanning approximately 16,000 × 16,000 pixels with a pixel resolution of roughly 12.5 cm × 12.5 cm. The images in the VEDAI dataset are available in two modalities: RGB and IR, representing the same scenes. The VEDAI dataset comprises 1,246 images that spotlight a variety of settings, including grasslands, highways, mountainous regions, and urban landscapes. These images are resized to either 1024 × 1024 or 512 × 512 pixels for analysis. The objective of the dataset is to identify 11 distinct classes of vehicles, encompassing categories such as cars, pickups, campers, and trucks.

3.4 Assessment Indicators

The accuracy assessment measures the agreements and differences between the detection result and the reference. The recall, precision, and mean Average Precision (mAP) are used as accuracy metrics to evaluate the performance of the

Table 1: Single modality and multi-modalities comparison. We use class-wise average precision AP and mean average precision mAP_{50} as metrics.

	Car	Pickup	Camping	Truck	Other	Tractor	Boat	Van	mAP_{50} ↑
Single-Mode (RGB)	89.31	87.31	69.71	83.34	61.04	**80.08**	54.17	75.52	75.06
Single-Mode (IR)	87.84	86.00	78.54	72.09	45.44	59.95	55.02	**81.77**	70.83
Multi-Modal (RGB+IR)	**90.80**	**87.79**	**83.31**	**83.83**	**69.48**	78.78	**85.23**	80.11	**82.42**

Fig. 2: Visual results of infrared single mode detection, visible light single mode detection and infrared and visible light multi-mode fusion detection.

methods. The calculations of the precision Pr and recall Re metrics are defined as

$$\mathrm{Pr} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \ \mathrm{Re} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{11}$$

where the true positive (TP) and true negative (TN) denote correct prediction, and the false positive (FP) and false negative (FN) denote incorrect outcome. The precision and recall are correlated with the commission and omission errors, respectively. The mAP is a comprehensive indicator obtained by averaging AP values, which uses an integral method to calculate the area enclosed by the precision–recall curve and coordinate axis of all categories. Hence, the mAP can be calculated by

$$\mathrm{mAP} = \frac{\mathrm{AP}}{N} = \frac{\int_0^1 \mathrm{Pr}(\mathrm{Re}) d\mathrm{Re}}{N}, \tag{12}$$

where N is the number of categories. Model complexity and computational cost are quantified using metrics of Giga Floating-point Operations Per Second (GFLOPs) and parameter size.

Table 2: Comparative Experiments. We use class-wise average precision AP, mean average precision mAP_{50}, Parameters computation and GFLOPS as metrics.

Method	Car	Pickup	Camping	Truck	Other	Tractor	Boat	Van	mAP_{50} ↑	Params(M) ↓	GFLOPs(G) ↓
YOLOv3 [18]	84.57	72.68	67.13	61.96	43.04	65.24	37.10	58.29	61.26	61.54	49.68
YOLOv3tiny [18]	-	-	-	-	-	-	-	-	58.10	8.70	13.00
YOLOv4 [4]	85.46	72.84	72.38	62.82	48.94	68.99	34.28	54.66	62.55	52.51	38.23
YOLOv5s [19]	80.81	68.48	69.06	54.71	46.79	64.29	24.25	45.96	56.79	7.07	**5.32**
YOLOv5m [19]	82.53	72.32	68.41	59.25	46.20	66.23	33.51	57.11	60.69	21.07	16.24
YOLOv5l [19]	82.83	72.32	69.92	63.94	48.48	63.07	40.12	56.46	62.16	46.64	36.70
YOLOv5x [19]	84.33	72.95	70.09	61.15	49.94	67.35	38.71	56.65	62.65	87.25	69.71
YOLOv5n [19]	-	-	-	-	-	-	-	-	59.30	7.00	16.00
YOLOv7 [23]	-	-	-	-	-	-	-	-	72.20	36.90	104.70
YOLOv8s [9]	-	-	-	-	-	-	-	-	65.60	11.16	28.60
YOLOv8m [9]	-	-	-	-	-	-	-	-	72.60	25.90	79.10
YOLOv8l [9]	-	-	-	-	-	-	-	-	74.80	43.60	165.40
YOLOv8x [9]	-	-	-	-	-	-	-	-	76.90	68.20	258.10
SuperYOLO [28]	**91.13**	85.66	79.30	70.18	57.33	**80.41**	60.24	76.50	75.09	**4.85**	17.98
ESM-YOLO	90.80	**87.79**	**83.31**	**83.83**	**69.48**	78.78	**85.23**	**80.11**	**82.42**	80.18	65.00

4 Experimental Results

4.1 Experimental Platform and Parameter Settings

To evaluate the effectiveness of the ESM-YOLO model, a comparison test and an ablation test were designed. The experiments were conducted on a machine equipped with an AMD7945 CPU and an NVIDIA RTX4060 GPU. The standard stochastic gradient descent (SGD) [6] is used to train the network with a momentum of 0.937, a weight decay of 0.0005 for the Nesterov accelerated gradients utilized, and a batch size of 2. The learning rate is set to 0.01 initially. The entire training process involves 300 epochs. The input image size is 1024× 1024 for training and 512×512 for testing. For loss function, $\alpha_o^a = 1.0, \alpha_l^a = 0.05, \alpha_c^a = 0.5, a = 0, 1, 2$, and $\lambda_o = \lambda_l = \lambda_c = 1.0$. The parameters $p^{RGB}$and p^{IR}in Eq. (1) are set to 0.5.

4.2 Comparison Algorithms and Results

To evaluate the usability and effectiveness of the ESM-YOLO model for detecting small targets in multi-modal remote sensing application scenarios, a comparative study was conducted with a variety of algorithms under the same conditions.

These results underscore the high detection accuracy of the ESM-YOLO model in recognizing small targets through the fusion of infrared and visible multimodal data. Notably, ESM-YOLO excels in the recognition of boats and vans. Comprehensive experimental data are presented in Tab. 2. Qualitative results of the different algorithms are shown in Fig. 3. The improved algorithm achieves a significant improvement in the accuracy of small target detection in remote sensing. This advancement in detection accuracy and reduction in missed detections is attributed to the optimization of feature extraction, fusion, and representation capabilities. Fig. 4 presents the Precision-Recall curves for a multi-class object detection model, evaluated across various classes of vehicles

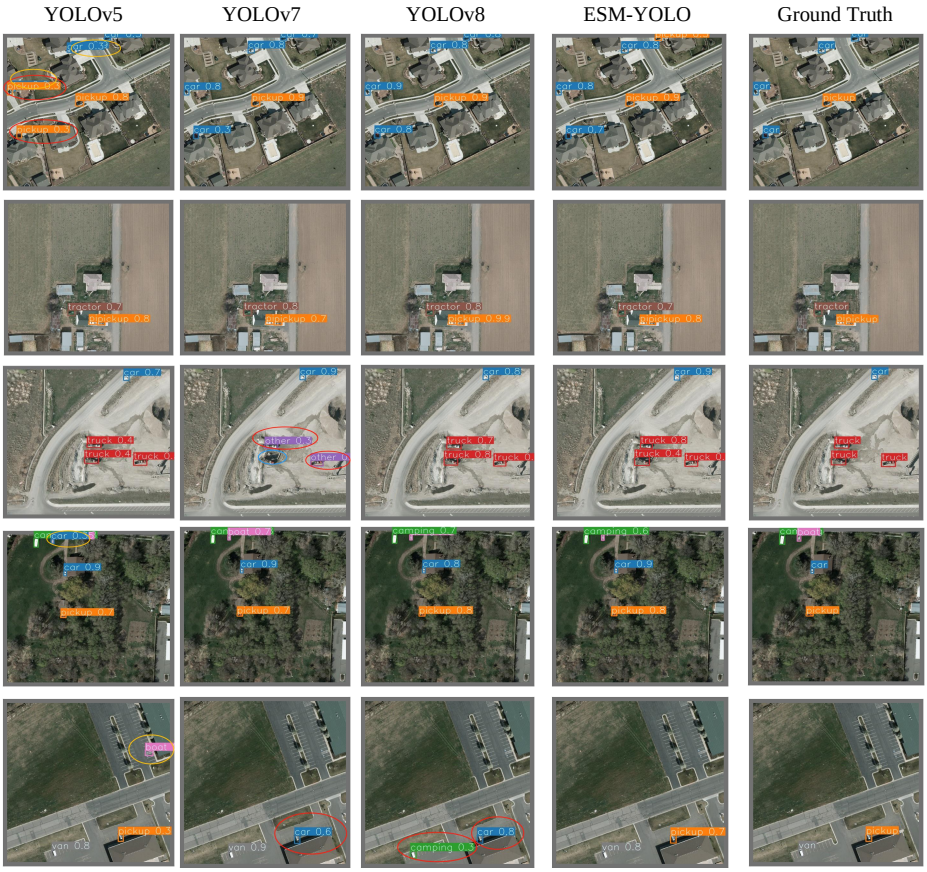

Fig. 3: Results of various algorithms are presented through visualization. The red cycles represent the false alarms, the yellow ones denote the FP detection results, and the blue ones are FN detection results.

within aerial imagery datasets. The model's performance is quantified through the mean Average Precision (mAP) metric at an Intersection over Union (IoU) threshold of 0.5, which serves as a balanced measure of precision and recall for each class.

4.3 Influence of Multi-modalities

To validate the substantial enhancement in recognition accuracy for small target detection achieved by fused multi-modal data over unimodal data post-network learning, a comparative analysis is conducted under identical conditions. This analysis encompasses the detection outcomes of unimodal visible, unimodal infrared, and multi-modal fusion-based small target recognition. Specifically, the mean average precision (mAP) for the unimodal visible model is 75.06%, while

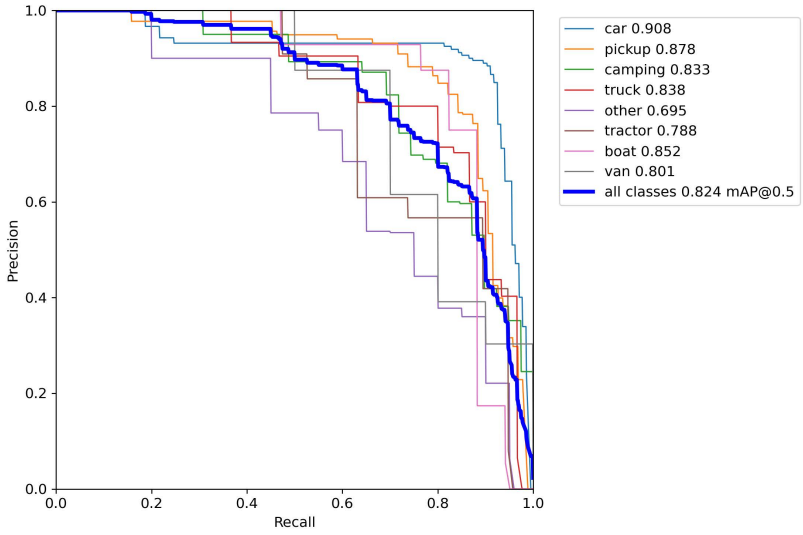

Fig. 4: ESM-YOLO Model P-R Curve Plot.

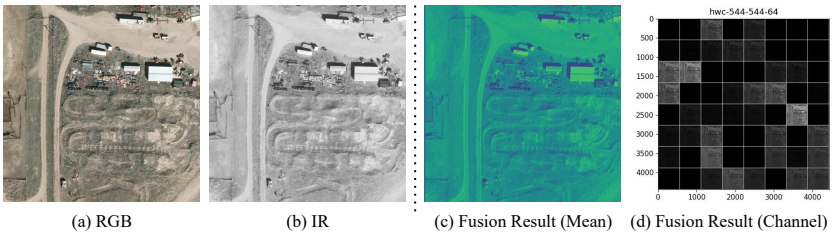

Fig. 5: Visualization of BEF fusion results. (a) RGB input. (b) IR input. (c) and (d) are fusion results indicated by the mean or channel.

that of the unimodal infrared model stands at 70.83%. Notably, the multi-modal ESM-YOLO model demonstrates a notable improvement in small target detection accuracy. The comprehensive experimental results are presented in Table Tab. 1.

Fig. 2 presents a comparative visual analysis of detection outcomes between single modality and multi-modality detection methods, highlighting their respective efficacies in object detection. The ESM-YOLO model, which employs multi-modal fusion, demonstrates a significant reduction in both false positives and false negatives, thereby alleviating the elevated false alarm rates that are often inherent to single modality detections. This comparative visualization clearly delineates the superiority of the ESM-YOLO model in enhancing detection accuracy and reliability.

Table 3: Ablation experiments on Bilateral Excitation Fusion (BEF), Compact BottleneckCSP (CBCSP), and Improved Atrous Spatial Pyramid Pooling (IASPP) .

No.	Basic backbone	BEF	CBCSP	IASPP	mAP_{50} ↑
1	✓	-	-	-	62.65
2	✓	✓	-	-	75.09
3	✓	✓	✓	-	81.07
4	✓	✓	✓	✓	**82.42**

4.4 Ablation Experiments

To validate the effectiveness of the improved model proposed in this paper, a series of ablation experiments were conducted on the VEDAI dataset to assess the impact of each module on the improved model. The results of the experiments are shown in Tab. 3. As can be seen from Tab. 3, the detection accuracy of the model after optimizing the network structure is significantly improved over the original model under the same settings. The baseline model establishes an initial mean Average Precision (mAP) of 62.65%.

Impact of Bilateral Excitation Fusion The incorporation of the Bilateral Excitation Fusion (BEF) module significantly enhances the mAP to 75.09%, highlighting its substantial contribution to improving detection accuracy. We also present a visualization of fusion results of our BEF in Fig. 5. Note that Fig. 5(c) is obtained by averaging along the channel axis and Fig. 5(d) represents all channels. It can be observed that our fusion block can extract comprehensive features of the two modalities, which contributes to downstream tasks.

Impact of Compact BottleneckCSP Further augmentation is observed with the integration of the Compact BottleneckCSP (CBCSP) structure, which elevates the mAP to 81.07%. This increment underscores the effectiveness of CBCSP in extracting features for small target detection.

Impact of Improved Atrous Spatial Pyramid Pooling In this study, we present an optimal configuration that integrates the Improved Atrous Spatial Pyramid Pooling (IASPP) structure alongside several other enhancements, achieving a remarkable mean Average Precision (mAP) of 82.42%. This result underscores the synergistic effects of the integrated modules in enhancing the detection capabilities of the model. An ablation study conducted in this research emphasizes the crucial role each component plays in boosting the model's overall performance. Collectively, our findings demonstrate the robustness of the ESM-YOLO model and its capability in achieving precise small target detection within multi-modal remote sensing applications.

Furthermore, we conducted an in-depth analysis of the performance of the IASPP and CBCSP modules under various configurations, with a particular focus on the sensitivity of key parameters. Specifically, within the IASPP module,

we evaluated the impact of various activation functions, including SELU [16], Gelu [11], and ReLU [3] *etc.*, on the model's accuracy. Experimental data verified significant differences in performance based on the choice of activation function. Similarly, we conducted a parallel analysis of the activation functions within the CBCSP module, further corroborating our initial findings. These fine-grained analyses provide a comprehensive understanding of the model's behavior and offer valuable insights for future optimization efforts. Refer to Tab. 4 for a detailed overview of our experimental results.

Table 4: Effect of Different Activation Functions on Recognition Accuracy (mAP ↑) of IASPP Modules.

Methods	Mish [14]	SELU [16]	Hard_sigmoid [20]	Gelu [11]	ReLU [3]	SiLU [8]	Ours
mAP ↑	77.84	75.21	77.88	79.31	79.96	80.18	**80.68**

5 Conclusions

This paper introduces an advanced algorithm, ESM-YOLO, designed for the detection of small, multi-scale targets within multi-modal remote sensing data. The algorithm incorporates a Bilateral Excitation Fusion (BEF) module to bolster the network's multi-modal fusion capabilities and promote more effective cross-modal learning. Furthermore, the Improved Atrous Spatial Pyramid Pooling (IASPP) structure has been designed to enhance the network's feature extraction capabilities. Additionally, the integration of a Compact BottleneckCSP (CBCSP) in the Head section, specifically optimized for small target detection, has led to a significant enhancement in the model's detection performance. The experimental outcomes demonstrate that ESM-YOLO achieves a 16.82% improvement in mean Average Precision (mAP) over YOLOv8s, a 7.62% improvement over YOLOv8l, and a 5.52% improvement over YOLOv8x. When compared to YOLOv5x, ESM-YOLO shows a 19.77% increase in mAP at an IoU of 0.5, a reduction of 7.0696 million parameters, and a decrease of 4.71 in GFLOPs. These results indicate that the ESM-YOLO model is more accurate in multi-modal fusion for small target detection, offering practical benefits for real-world applications.

6 Acknowledgements

This work was supported by the Chinese Academy of Sciences Key Laboratory Fund - Space-Based Intelligent Electromagnetic Spectrum Monitoring (E32213A01S).

References

1. An, C., Wang, Y., Zhang, J., Nguyen, T.Q.: Self-supervised rigid registration for multimodal retinal images. IEEE Transactions on Image Processing **31**, 5733–5747 (2022) 4
2. Bai, Y., Zhang, Y., Ding, M., Ghanem, B.: Sod-mtgan: Small object detection via multi-task generative adversarial network. In: Proceedings of the European conference on computer vision (ECCV). pp. 206–221 (2018) 4
3. Bai, Y.: Relu-function and derived function review. In: SHS Web of Conferences. vol. 144, p. 02006. EDP Sciences (2022) 14
4. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020) 10
5. Bogdoll, D., Nitsche, M., Zöllner, J.M.: Anomaly detection in autonomous driving: A survey. In: CVPR. pp. 4488–4499 (2022) 4
6. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers. pp. 177–186. Springer (2010) 10
7. Cheng, G., Yuan, X., Yao, X., Yan, K., Zeng, Q., Xie, X., Han, J.: Towards large-scale small object detection: Survey and benchmarks. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) 4
8. Elfwing, S., Uchibe, E., Doya, K.: Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. Neural networks **107**, 3–11 (2018) 14
9. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics yolo: Software for object detection. version 8.0.0. https://github.com/ultralytics/ultralytics (2023), accessed July 3, 2024 10
10. Kang, X., Yin, H., Duan, P.: Global–local feature fusion network for visible–infrared vehicle detection. IEEE Geoscience and Remote Sensing Letters **21**, 1–5 (2024) 4
11. Lee, M.: Gelu activation function in deep learning: a comprehensive mathematical analysis and performance. arXiv preprint arXiv:2305.12073 (2023) 14
12. Li, J., Liang, X., Wei, Y., Xu, T., Feng, J., Yan, S.: Perceptual generative adversarial networks for small object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1222–1230 (2017) 4
13. Liu, H., Ye, Y., Zhang, J., Yang, C., Zhao, Y.: Comparative analysis of pixel level fusion algorithms in high resolution sar and optical image fusion. In: IGARSS. pp. 2829–2832. IEEE (2022) 3
14. Misra, D.: Mish: A self regularized non-monotonic activation function. arXiv preprint arXiv:1908.08681 (2019) 14
15. Noh, J., Bae, W., Lee, W., Seo, J., Kim, G.: Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9725–9734 (2019) 4
16. Rasamoelina, A.D., Adjailia, F., Sinčák, P.: A review of activation function for artificial neural network. In: 2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI). pp. 281–286. IEEE (2020) 14
17. Razakarivony, S., Jurie, F.: Vehicle detection in aerial imagery: A small target detection benchmark. Journal of Visual Communication and Image Representation **34**, 187–203 (2016) 5, 8

18. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018) 10
19. Ultralytics: Yolov5 homepage. https://github.com/ultralytics/yolov5 (2021), accessed July 3, 2024 7, 8, 10
20. Valmiki, G.C., Tirupathi, A.S.: Performance analysis between combinations of optimization algorithms and activation functions used in multi-layer perceptron neural networks (2020) 14
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 4
22. Vögtli, M., Sierro, L., Kneubühler, M., Schreiner, S., Gross, W., Queck, F., Kuester, J., Mispelhorn, J., Middelmann, W.: Hyperthun'22: A multi-sensor multi-temporal camouflage detection campaign. In: IGARSS. pp. 2153–2156. IEEE (2023) 1
23. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7464–7475 (2023) 10
24. Wang, C.Y., Liao, H.Y.M., Wu, Y.H., Chen, P.Y., Hsieh, J.W., Yeh, I.H.: Cspnet: A new backbone that can enhance learning capability of cnn. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 390–391 (2020) 4
25. Xiao, F., Tong, L., Wen, J., Wang, Y.: Decision-level fusion for road network extraction from sar and optical remote sensing images. In: IGARSS. pp. 7427–7430. IEEE (2023) 3
26. Xin, D., Xu, L., Chen, H., Yang, X., Zhang, R.: A vehicle target detection method based on feature level fusion of infrared and visible light image. In: 2022 34th Chinese Control and Decision Conference (CCDC). pp. 469–474. IEEE (2022) 3
27. Yu, B., Chen, W., Wang, W.: Research on industrial non-destructive testing technology based on improved yolov5s. In: ICTech. pp. 435–440. IEEE (2023) 4, 5
28. Zhang, J., Lei, J., Xie, W., Fang, Z., Li, Y., Du, Q.: Superyolo: Super resolution assisted object detection in multimodal remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–15 (2023) 2, 10
29. Zhang, Q., Zhou, L., An, J.: Real-time recognition algorithm of small target for uav infrared detection. Sensors **24**(10), 3075 (2024) 1
30. Zhang, T., Xie, G., Xie, X., Li, L., Zhang, H., Liang, Y.: An efficient bidirectional weighted global feature extraction algorithm for steel surface defects. In: 2023 China Automation Congress (CAC). pp. 6370–6375. IEEE (2023) 4
31. Zhou, J., Zhang, R., Zhao, W., Shen, S., Wang, N.: Aps-net: An adaptive point set network for optical remote-sensing object detection. IEEE Geoscience and Remote Sensing Letters **20**, 1–5 (2022) 1