TAF-YOLO: A Small-Object Detection Network for UAV Aerial Imagery via Visible and Infrared Adaptive Fusion

Zhanhong Zhuo , Ruitao Lu * , Yongxiang Yao , Siyu Wang , Zhi Zheng , Jing Zhang , Xiaogang Yang

Posted Date: 6 November 2025

doi: 10.20944/preprints202511.0375.v1

Keywords: small-object detection; multimodal; feature fusion; UAV aerial images; visible and infrared; YOLO

Article

TAF-YOLO: A Small-Object Detection Network for UAV Aerial Imagery via Visible and Infrared Adaptive Fusion

Zhanhong Zhuo [1], Ruitao Lu [1,*], Yongxiang Yao [2], Siyu Wang [3], Zhi Zheng [4], Jing Zhang [5] and Xiaogang Yang [1]

[1] College of Missile Engineering, Rocket Force University of Engineering, Xi'an 710025, China
[2] School of Remote Sensing Information Engineering, Wuhan University, Wuhan 430079, China
[3] School of Electrical and Control Engineering, Xuzhou University of Technology, Xuzhou 221018, China
[4] Department of Geography and Resource Management, The Chinese University of Hong Kong, Hong Kong
[5] State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China
* Correspondence: lrt19880220@163.com

Highlights

What are the main findings?

- An early-fusion multimodal small-object detection framework, TAF-YOLO, is proposed for UAV aerial imagery. It effectively integrates complementary visible and infrared information at the pixel level, thereby improving small-object detection accuracy while reducing missed detections and false alarms.
- TAF-YOLO removes the PANet from the baseline model, reduces redundant information and introduces lightweight DSAB into a multimodal detection framework, achieving efficient and accurate detection of small objects.

What are the implications of the main findings?

- The proposed framework shows that early-fusion of visible and infrared modalities improves robustness and precision in UAV-based small-object detection, suggesting a promising direction for multimodal sensing in complex environments.
- The integration of lightweight modules and removal of redundant network structures provide evidence that high detection accuracy can be achieved with reduced computational cost, supporting broader deployment in real-time and resource-constrained scenarios.

Abstract

Detecting small objects from UAV-captured aerial imagery is a critical yet challenging task, hindered by factors such as small object size, complex backgrounds, and subtle inter-class differences. Single-modal methods lack the robustness for all-weather operation, while existing multimodal solutions are often too computationally expensive for deployment on resource-constrained UAVs. To this end, we propose TAF-YOLO, a lightweight and efficient multimodal detection framework designed to balance accuracy and efficiency. First, an early fusion module, the Two-branch Adaptive Fusion Network (TAFNet), which adaptively integrates visible and infrared information at both pixel and channel levels before the feature extractor, maximizing complementary data while minimizing redundancy. Second, a Large Adaptive Selective Kernel (LASK) module that dynamically expands the receptive field using multi-scale convolutions and spatial attention, preserving crucial details of small objects during downsampling. Finally, an optimized feature neck architecture that replaces PANet's bidirectional path with a more efficient top-down pathway. This is enhanced by a Dual-Stream Attention Bridge (DSAB) that injects high-level semantics into low-level features, improving localization without significant computational overhead. On the VEDAI benchmark, TAF-YOLO achieves 67.2% mAP_{50}, outperforming the CFT model by 2.7% and demonstrating superior

 doi:10.20944/preprints202511.0375.v1

performance against seven other YOLO variants. Our work presents a practical and powerful solution that enables real-time, all-weather object detection on resource-constrained UAVs.

Keywords: small-object detection; multimodal; feature fusion; UAV aerial images; visible and infrared; YOLO

1. Introduction

With the rapid development of unmanned aerial vehicle (UAV) technology [1] and the miniaturization of onboard sensors [2], UAV-based visual perception has become a key tool in applications such as disaster monitoring [3], agricultural management [4], and urban surveillance [5]. Compared with satellite or ground-based observation, UAV imagery offers flexible viewpoints and high temporal resolution but also presents unique challenges for object detection: targets are typically small, densely distributed, and surrounded by complex backgrounds [6]. Furthermore, variations in flight altitude and environmental conditions result in diverse object scales and appearances, significantly increasing detection difficulty.

Although recent advances in deep learning, particularly in Convolutional Neural Networks (CNNs) and Transformer-based architectures, have led to remarkable progress in UAV remote sensing detection [7–10], methods relying solely on visible images are sensitive to challenging conditions such as low light, heavy fog, or rain, where their performance degrades sharply. In contrast, infrared (IR) imaging, which captures thermal radiation from objects, is robust to lighting variations and capable of all-weather operation, effectively highlighting targets obscured by poor illumination or occlusion. However, IR imagery inherently suffers from low spatial resolution and a lack of color and texture details. Therefore, visible and infrared modalities are naturally complementary: the former provides high-frequency texture and color information, while the latter contributes stable, illumination-invariant thermal cues.

Modern UAVs are often equipped with dual-sensor gimbals integrating both visible and infrared cameras, providing an ideal foundation for multimodal fusion (MF). Effectively fusing these two modalities can significantly enhance the robustness and reliability of UAV perception systems, which is crucial for improving the detection of small objects in complex environments. Consequently, building an efficient end-to-end visible-infrared fusion detection network that fully leverages inter-modality complementarity is an inevitable requirement for the advancement of intelligent visual perception technology.

Currently, deep learning-based visible-infrared fusion (VIF) methods [11] are primarily categorized into early (pixel-level), intermediate (feature-level), and late (decision-level) fusion. Early fusion [12], though computationally efficient, often relies on simple operations like addition or concatenation, failing to adequately model the deeper relationships between modalities. Intermediate [13] and late fusion [14] strategies (e.g., those using Transformer-based cross-attention [15,16]) better capture cross-modal interactions but still face challenges such as feature space misalignment, information loss, and insufficient adaptation to UAV-specific scenarios. Existing fusion frameworks generally suffer from three key limitations: overly simplistic early fusion strategies, information loss in later fusion stages, and inadequate adaptation to UAV small-object scenarios, where feature degradation is particularly severe.

To address these issues, this paper proposes TAF-YOLO, a multimodal small-object detection network designed specifically for UAV aerial imagery. Before feature extraction, TAFNet adaptively fuses visible and infrared information at both the pixel and channel levels. This early fusion strategy effectively integrates complementary information while suppressing redundant features, providing a high-quality input for subsequent detection. To mitigate feature degradation during downsampling, the LASK module dynamically adjusts its receptive field using multi-scale depthwise convolutions and spatial attention. Combined with an anti-aliasing filtering strategy, it preserves edge information and enhances the feature representation of small objects. In the network's neck, we introduce the

 doi:10.20944/preprints202511.0375.v1

lightweight DSAB module for the first time. It injects high-level semantics into shallow features via semantic-guided foreground-background routing and multi-scale context aggregation, significantly improving localization precision for small objects while maintaining semantic integrity.

Through these designs, TAF-YOLO achieves an excellent balance between detection accuracy and computational efficiency. The main contributions of this paper are summarized as follows:

1. We propose TAF-YOLO, a novel end-to-end multimodal fusion framework specifically engineered for robust small-object detection in complex UAV scenarios. By integrating a series of targeted optimizations, the framework establishes a SOTA balance between high detection accuracy and UAV-friendly computational efficiency.

2. To overcome key bottlenecks in multimodal feature fusion and propagation, we introduce a set of innovative modules. An efficient early fusion network, TAFNet, to maximize modal complementarity while preserving critical details at the input stage. The LASK module in the backbone to mitigate feature degradation of small objects via adaptive receptive fields and an anti-aliasing strategy. The DSAB module in the neck to enhance localization precision through targeted cross-layer semantic injection.

3. We conducted extensive experiments on public UAV visible-infrared datasets. The results demonstrate that, compared to current state-of-the-art methods, TAF-YOLO achieves superior performance in both accuracy and efficiency.

The rest of this article is arranged as follows. The related works are briefly reviewed in Section 2. The details of the proposed method are described in Section 3. Experiments and evaluations are provided in Section 4. A conclusion about this article is given in Section 5.

2. Related Work

2.1. Deep Learning Based Object Detection

Object detection aims to automatically identify and locate specific objects in spatial images. With the development of deep neural networks, deep learning based object detection technology has been widely applied on intelligent platforms such as UAVs. Currently, the predominant deep learning detection algorithms can be categorized into the CNN-based two-stage detectors, CNN-based one-stage detectors, and the more recent Transformer-based detectors.

Two-stage detectors, represented by R-CNN [17], Fast R-CNN [18], Faster R-CNN [19], and Cascade R-CNN [20], first generate region proposals and then perform classification and regression in separate steps to complete detection. These approaches offer high accuracy but are less suitable for embedded or UAV platforms due to the high computational demands. In contrast, one-stage detectors such as SSD [21] and RetinaNet [22] leverage dense anchor priors or focal loss to achieve a significant speed boost while maintaining accuracy. Meanwhile, Transformer-based detectors, such as the DETR family [23,24], reformulate object detection as a set prediction problem, relying on global self-attention mechanisms and a Hungarian algorithm-based matching strategy to realize an end-to-end framework that obviates the need for non-maximum suppression (NMS). Furthermore, models such as Swin Transformer [25] and ViT [26] have been extensively employed as backbones or detection heads, integrating long-range dependency modeling with hierarchical feature representation to enhance both accuracy and scalability.

In one-stage detection methods, many studies have adopted the YOLO series as the foundational framework for UAV-based small-object detection owing to its efficient detection capability. From the original YOLO [27] to the widely deployed YOLOv5 [28] and YOLOv8 [29] models in both academia and industry, the series has undergone continuous evolution. Recent variants—such as YOLOv10 [30], which emphasizes NMS-free operation and high accuracy, as well as the lightweight YOLOv11 [31] and the attention-based YOLOv12 [32]—further refine the architecture and extend its applicability across diverse real-world scenarios.

Small objects in UAV aerial imagery are densely distributed and complex in background, causing the traditional detection algorithms to be prone to false and missed. Wu et al. [33] proposed

 doi:10.20944/preprints202511.0375.v1

MV-YOLO which leverages the Mamba Self-Attention Vision Module (MSAVM) to enhance long-range interactions, combined with the Context Clue Guided Module (CCGM) to refine the representation of small objects, thereby achieving efficient and accurate detection of small objects. Liu et al. [34] proposed MFAE-YOLO, which introduces a Global Feature Fusion Processing (GFFP) module to enhance global feature extraction and an Accurate Intersection over Union (AIoU) loss function to improve the precision of bounding box regression, thereby improving the detection accuracy of small objects in densely packed scenarios. Cao et al. [35] proposed GCL-YOLO, a lightweight YOLO network based on GhostConv, which introduces a new prediction head special for UAV small objects and uses the Focal-Efficient Intersection over Union (Focal-EIOU) loss as the localization loss function, improving detection accuracy while being lightweight. Detecting small objects in infrared images, especially in low-contrast and complex backgrounds, remains challenging. Tang et al. [36] proposed IRSTD-YOLO, an infrared small target detection network, which introduces the infrared small target enhancement (IRSTE) module and the edge and feature extraction (EFE) module, enhancing small target representation and detection robustness. Hao et al. [37] proposed YOLO-SR, which combines image super-resolution technology with YOLO, leveraging the structure-preserving super-resolution (SPSR) to preprocess the input infrared images before they enter the detection network, which enhances the accuracy of the network.

Although the above methods have achieved significant improvements in detection accuracy and speed, single-modal detectors remain vulnerable to external factors such as lighting, weather variations, and occlusion. The limited representational capacity of a single modality makes it difficult to maintain consistent and reliable detection performance under complex environmental conditions. In contrast, multimodal detection techniques provide an effective solution by integrating complementary information from multiple sources.

2.2. Multimodal Fusion Object Detection

With the development of sensor technology, multimodal data has been widely applied in fields such as autonomous driving [38] and remote sensing [39]. Multimodal object detection addresses the detection failure problem caused by light change, occlusion or noise in a single mode by jointly optimizing the datas from different physical sensors such as optical and infrared sensors. It fully leverages the complementary characteristics of different modal datas to effectively improve the robustness and accuracy of detection. Different sensors in multimodal object detection each have their own advantages but also exhibit inherent limitations. Visible-light and infrared cameras are particularly suitable for UAV platforms due to their compact size, low weight, low power consumption, and real-time imaging capabilities. Moreover, these two modalities are highly complementary in terms of information: visible-light images provide rich texture and structural details, while infrared images offer strong robustness under low-light and complex background conditions.Therefore, the VIF has become the mainstream approach in current research on multimodal small-object detection for UAVs.

In object detection tasks, features at different levels play complementary roles in determining the final detection performance. Low-level features typically contain rich local texture, edge, and spatial location information, while high-level features encode more global contextual and semantic information. The ability to effectively integrate these features at appropriate stages directly impacts the perception and detection capabilities of detector. Currently, there are three fusion paradigms: early fusion, intermediate fusion, and late fusion.

Intermediate fusion generally refers to the process of extracting features separately from the original data of different modalities and integrating them by leveraging the complementary advantages of distinct feature maps. This approach aims to achieve a balance between fine-grained local details and broader semantic context, thereby enhancing the overall representation capability of the model. Wang et al. [40] proposed a cross-scale dynamic convolution-driven YOLO fusion network, which introduces a novel cross-scale dynamic convolution fusion module to adaptively extract and fuse bimodal features concerning on data distribution. Shen et al. [41] proposed a novel

 doi:10.20944/preprints202511.0375.v1

feature fusion framework of dual cross-attention transformers, which utilizes an iterative interaction mechanism to share parameters among block-wise multimodal transformers, reducing model complexity and computational cost. Zeng et al. [42] proposed MMI-Det, a multi-modal fusion method for visible and infrared object detection, which introduces the fusion focus module and the information guided module, thereby enhancing the ability to perceive and combine visible-infrared modal information. Intermediate fusion effectively combines complementary features of different modalities during the later stage of feature extraction, effectively utilizing local and global information. Intermediate fusion integrates multimodal information at the feature level, effectively capturing complementary representations and balancing local detail with global semantics. However, since features are extracted independently before integration, this strategy often leads to the loss of modality-specific information and incomplete preservation of intrinsic feature relationships.

Late fusion improves detection accuracy by integrating the predictions of separate modality networks. This preserves the distinctive features of each modality and reduces cross-modal interference. Tang et al. [43] proposed a novel decision fusion module (DFM) for making an intermediate decision based on the features in the current layer and then fuse its results with the original features before passing them to the next layers to avoid information loss during feature abstraction and fully utilize lower layer features.

In contrast, early fusion combines visible and infrared images at the input level, preserving complementary texture, color, and thermal information. This approach enables more complete feature extraction and reduces information loss, while maintaining a simpler and more efficient network structure. Zhang et al. [44] proposed SuperYOLO, which utilizes a symmetric compact MF to extract supplementary information from various data and the assisted super resolution (SR) learning, showing a favorable accuracy–speed tradeoff. Chen et al. [45] proposed a MF framework—Fused Modality Enhancement Network, which consists of two fusion modules: the Common Feature Enhancement Module and the Difference Feature Enhancement Module. This method achieves accurate detection of low light objects at night while reducing the computational complexity of the model.

Considering the above factors, this study adopts an early fusion strategy. On resource-constrained UAV platforms, early fusion offers a lightweight and efficient solution by requiring only a single feature extraction pathway, thus reducing computational and memory overhead. Meanwhile, by integrating visible and infrared data at the input level, this approach fully exploits their complementary characteristics.

3. Methodology

3.1. Overview of Our Method

The comprehensive framework of our propose TAF-YOLO network, designed for multimodal small-object detection in UAV aerial imagery, is illustrated in Figure **1**.

First, we explore different MF methods and propose a TAFNet. By adaptively fusing visible and infrared modality images at both the pixel and channel levels before the backbone network, we achieved full integration of complementary features while preserving fine-grained spatial details that are crucial for small objects. Secondly, the proposed LASK module replaces the standard convolutional layers at the P3 and P4 stages in the YOLOv12 backbone to alleviate feature degradation during downsampling. By employing multi-scale depthwise convolutions and adaptive spatial attention, it selectively emphasizes informative regions and integrates an anti-aliasing filter to preserve edge and texture information, thereby enhancing the representation of small objects. Finally, the feature path simplification strategy retains a pure top-down feature fusion pathway. On this basis, the DSAB introduces a lightweight semantic guidance strategy that injects high-level semantics into low-level features and performs multi-scale context aggregation. The top-down feature fusion pathway combined with the lightweight semantic guidance strategy improves the semantic

 doi:10.20944/preprints202511.0375.v1

completeness of small object features, thereby boosting detection accuracy and maintaining computational efficiency.

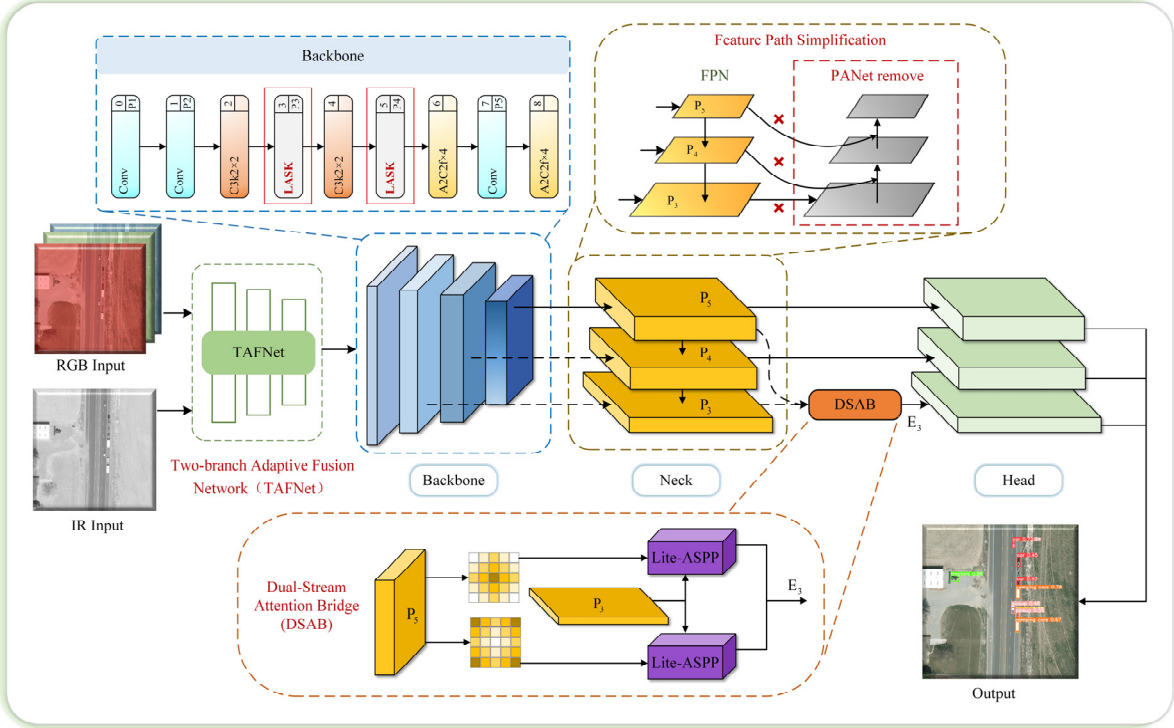

Figure 1. Overvall framework of the TAF-YOLO.

3.2. Two-branch Adaptive Fusion Network

The diversity of multimodal data can effectively enhance informational complementarity, thereby significantly improving the accuracy and perceptual capability of object detection models. In UAV object detection tasks, the method of feature fusion among different modalities greatly influences detection performance. If intermediate fusion is conducted during the feature extraction stage, it may lead to information loss or distortion, which is particularly detrimental to small-object detection in UAV imagery, as even the slightest detail loss can cause a marked decline in recognition accuracy. Similarly, although late fusion can integrate multimodal information at the output level, it only operates at the final stage, failing to fully exploit the complementary nature of low-level features across modalities, while also increasing computational complexity and processing time.

However, directly concatenating or summing these features often causes redundant information propagation and suboptimal modality weighting, as different regions may rely on different modalities. To address this issue, we propose a Two-branch Adaptive Fusion Network(TAFNet), which dynamically adjusts the contribution of each modality in a pixel-level and channel-level adaptive mechanism. The TAFNet enables the network to emphasize the most informative modality at each spatial location while suppressing redundant features, thereby achieving high-quality fusion with minimal computational cost. As shown in Figure 2, the input modalities of RGB and infrared are denoted as $I_{RGB} \in \mathbb{R}^{C_v \times H \times W}$, $I_{IR} \in \mathbb{R}^{C_i \times H \times W}$, respectively. To align both modalities into a common feature space, each branch is processed by a 3×3 convolution, batch normalization, and SiLU activation:

$$M_{RGB} = \phi(\mathcal{B}(\text{Conv1}(I_{RGB}))),\ M_{IR} = \phi(\mathcal{B}(\text{Conv2}(I_{IR}))) \tag{1}$$

where Conv1 and Conv2 represent the 3×3 convolution, $\mathcal{B}$ represents batch normalization and $\phi(\cdot)$ denotes the SiLU function.

 doi:10.20944/preprints202511.0375.v1

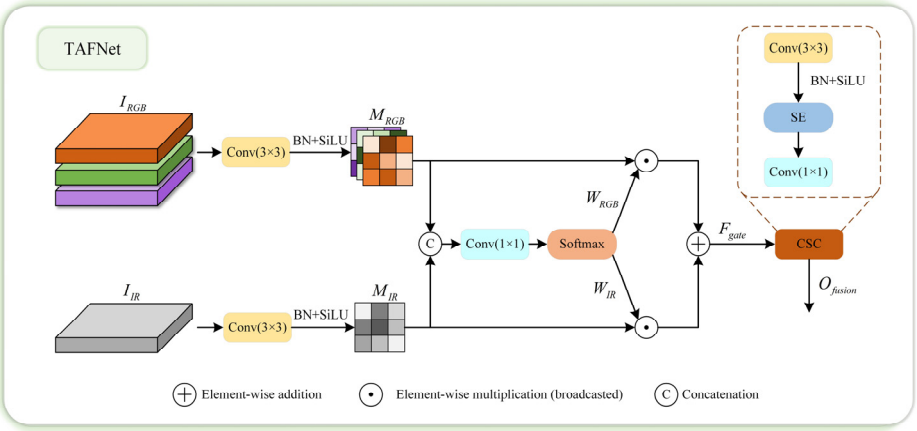

Figure 2. The Schematic of TAFNet.

After alignment, both features share the same intermediate dimensionality d, ensuring the two features have a consistent expression space before subsequent fusion. Then, The two aligned features are concatenated along the channel dimension and passed through a 1×1 convolution to generate gating logits:

$$A = \text{Conv3}(\text{Concat}(M_{RGB}, M_{IR})) \in \mathbb{R}^{2\times H\times W} \tag{2}$$

Here, $A = [A_{RGB}, A_{IR}]$ represents the unnormalized modality scores at each spatial location, Conv3 denotes the 1×1 convolution operations, and Concat represents the concatenation operation.

A modality-wise softmax is then applied along the modality dimension (channel axis) to obtain normalized pixel-wise weight maps $W_{RGB}, W_{IR} \in \mathbb{R}^{1\times H\times W}$:

$$W_{RGB} = \frac{e^{A_{RGB}}}{e^{A_{RGB}} + e^{A_{IR}}}, W_{IR} = \frac{e^{A_{IR}}}{e^{A_{RGB}} + e^{A_{IR}}} \tag{3}$$

which ensures $W_{RGB} + W_{IR} = 1$ for every pixel. The softmax operation compares modal scores at each spatial location and converts them into a probability-like allocation of attention between modalities, thus realizing pixel level mode selection. The fused feature is obtained as:

$$F_{gate} = W_{RGB} \odot F_{RGB} + W_{IR} \odot F_{IR} \tag{4}$$

where $\odot$ denotes element-wise multiplication with broadcasting.This soft gating strategy allows the network to adaptively select the most reliable modality based on spatial context, achieving dynamic weighted fusion.

To further enhance inter-channel dependencies and suppress noisy responses, we introduce a lightweight Squeeze-and-Excitation (SE) calibration after fusion. The final fused feature refined as:

$$F' = \phi(\mathcal{B}(\text{Conv4}(F_{gate}))) \tag{5}$$

$$O_{fusion} = \text{Conv5}\left(F' \otimes (\sigma(W_2 \text{ReLU}(W_1 GAP(F'))))\right) \tag{6}$$

where $\otimes$ represents the element-wise multiplication between the original feature map and the excitation weights, $GAP(\cdot)$ represents the global average pooling operation, $\sigma(\cdot)$ represents the Sigmoid function, W_1 and W_2 denote two fully connected layers forming a bottleneck structure, Conv4 and Conv5 represent the 3×3 and 1×1 convolution operations, respectively.

The unique advantage of the TAFNet lies in its ability to combine the spatial adaptivity of attention mechanisms with the efficiency of channel recalibration. Traditional fusion approaches often fail to adapt to local modality variations, leading to loss of fine-grained details crucial for small-object detection. In contrast, TAFNet dynamically generates pixel-wise modality weights through spatial Softmax gating, achieving adaptive feature fusion without redundant computation.

 doi:10.20944/preprints202511.0375.v1

Furthermore, the embedded SE structure enhances inter-channel dependency modeling, allowing subtle features from different modalities to be selectively amplified.

3.3. Large Adaptive Selective Kernel Module

As shown in Figure 3, the proposed backbone network is designed to extract both fine-grained texture features and high-level semantic information from UAV aerial images fused by TAFNet. It consists of three main types of modules: the C3k2 block, the LASK (Large Adaptive Selective Kernel) module, and the A2C2f structure. The C3k2 blocks in the shallow layers are responsible for capturing detailed spatial textures and low-level features, which are crucial for detecting small objects with weak visual cues.

To address the feature degradation caused by conventional downsampling, we replace the standard convolutional layers at the P3 and P4 stages with the proposed LASK module. This module dynamically adjusts its receptive field through multi-scale depth-wise convolutions and an adaptive spatial attention mechanism, enabling the network to selectively focus on informative regions while suppressing background noise. Moreover, the LASK module integrates an anti-aliasing filter that preserves structural details and reduces information loss during $\text{stride} = 2$ downsampling, which is particularly beneficial for small and densely distributed objects in UAV aerial images. In deeper stages, the A2C2f modules employ area attention and multi-branch feature fusion to enhance global contextual perception and semantic consistency, helping the model distinguish small objects from cluttered backgrounds.

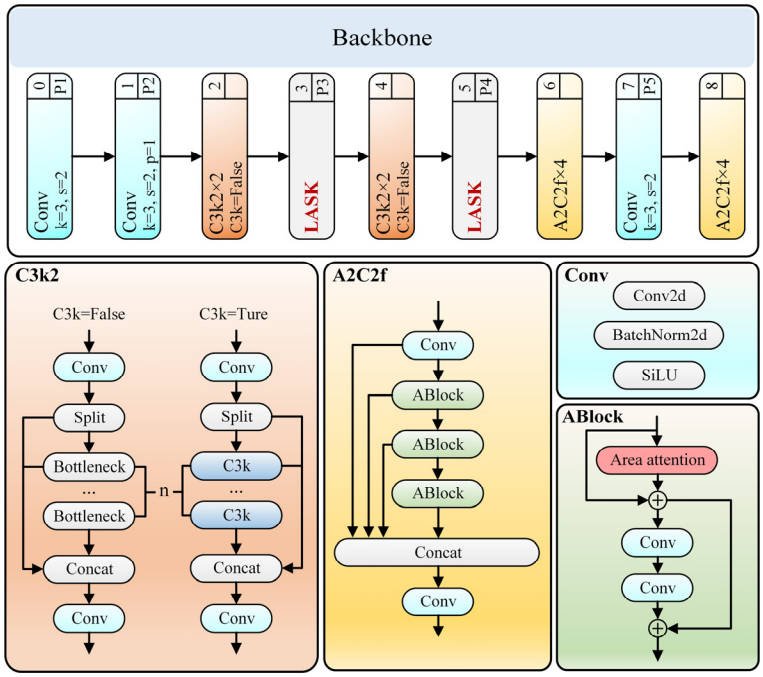

Figure 3. Backbone structure of TAF-YOLO.

The proposed LASK is illustrated in Figure 4, aiming to introduce a multi-scale selective convolution mechanism coupled with an anti-aliasing downsampling strategy. Unlike standard convolutional blocks, which use a single receptive field and straightforward $\text{stride} = 2$ downsampling, LASK performs multi-scale feature extraction, adaptive spatial weighting, and low-pass filtering before subsampling.

 doi:10.20944/preprints202511.0375.v1

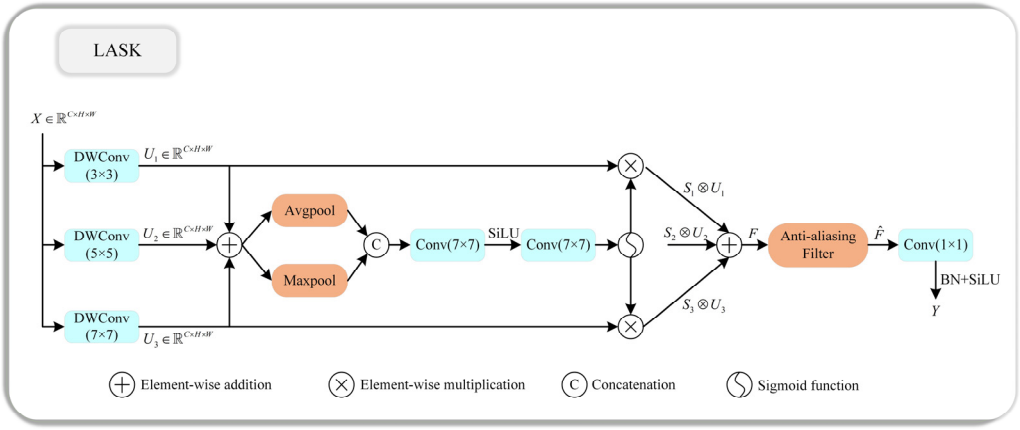

Figure 4. The illustration of the LASK module.

To capture both local and global spatial context, LASK employs multiple depth-wise convolutional branches with different receptive fields.Specifically, three depth-wise convolutions are applied to the same input feature map $X \in \mathbb{R}^{C\times H\times W}$:

$$U_1 = \mathcal{F}_{\mathrm{DW}}^{3\times 3}(X), U_2 = \mathcal{F}_{\mathrm{DW}}^{5\times 5}(X), U_3 = \mathcal{F}_{\mathrm{DW}}^{7\times 7, d=3}(X) \tag{7}$$

where $\mathcal{F}_{\mathrm{DW}}^{k\times k}(\cdot)$ represents a depth-wise convolution with kernel size k and dilation rate d .

The three branches respectively emphasize local texture(U_1), mid-range context(U_2), and long-range semantic structure(U_3). And $U_1, U_2, U_3 \in \mathbb{R}^{C\times H\times W}$, each maintaining the same number of channels C as the input. Because these branches are depth-wise and channel-aligned, each channel in U_i corresponds to the same semantic component across different receptive fields. In this setting, element-wise summation is the most natural and semantically consistent fusion strategy. The multi-scale feature maps are first aggregated through element-wise summation to form $\tilde{U} = U_1 + U_2 + U_3$. Unlike channel concatenation, which would artificially expand the dimensionality and require additional projection to realign the channels, element-wise addition preserves a one-to-one correspondence between feature channels across scales. This design ensures that each channel of U_i integrates multi-scale spatial responses that describe the same semantic entity under different receptive contexts. Then, channel-wise average pooling and max pooling are applied to capture spatial descriptors that reflect different types of activation statistics. These two pooled feature maps are concatenated and passed through a pair of large 7×7 convolutional layers to form a spatial attention tensor:

$$S = \sigma\left(\mathcal{F}^{7\times 7}(\mathrm{Concat}(P_{\mathrm{avg}}(\tilde{U}), P_{\max}(\tilde{U})))\right) \tag{8}$$

where $\sigma(\cdot)$ represents the Sigmoid function, $\mathcal{F}^{7\times 7}(\cdot)$ denotesa two-layer transformation composed of a 7×7 convolution, a SiLU nonlinearity, and another 7×7 convolution, $P_{\mathrm{avg}}(\cdot)$ and $P_{\max}(\cdot)$ represent channel-wise average pooling and max pooling operation, respectively.

The resulting attention map $S = [S_1, S_2, S_3] \in \mathbb{R}^{3\times H\times W}$ generates three spatial attention coefficients, each corresponding to one of the convolution branches.This mechanism adaptively assigns higher importance to the most informative regions at each receptive field scale. Each feature branch is then modulated by its corresponding attention mask, and all branches are fused to form the final aggregated feature:

$$F = \sum_{i=1}^{3}(S_i \otimes U_i) \tag{9}$$

where $\otimes$ denotes element-wise multiplication.This selective fusion process effectively allows the model to emphasize spatially salient features across multiple scales while suppressing redundant or

 doi:10.20944/preprints202511.0375.v1

noisy background activations.As a result, the fused feature map F carries richer semantic cues with enhanced discriminative power for object localization and recognition.

In conventional CNNs, $\text{stride} = 2$ convolutions can introduce aliasing artifacts that blur object boundaries and degrade high-frequency textures. To address this issue, LASK introduces an anti-aliasing filter $\mathcal{L}(\cdot)$ before downsampling.This filter is implemented as a depth-wise convolution using a fixed Gaussian-like kernel. The filtered feature $\hat{F}$ is obtained as $\hat{F} = \mathcal{L}(F) = F * K_{5\times5}$. The low-pass filter smooths high-frequency components while retaining important edge information.

$$K_{5\times5} = \frac{1}{256}\begin{bmatrix} 1 & 4 & 6 & 4 & 1 \\ 4 & 16 & 24 & 16 & 4 \\ 6 & 24 & 36 & 24 & 6 \\ 4 & 16 & 24 & 16 & 4 \\ 1 & 4 & 6 & 4 & 1 \end{bmatrix} \tag{10}$$

This anti-aliasing operation reduces spatial distortion during sub-sampling, leading to more stable feature representations in deeper layers. After anti-aliasing, a point-wise convolution (1×1) is applied to mix channel information and restore the desired output dimensionality. Subsequently, Batch Normalization and the SiLU activation function are employed to stabilize training and introduce non-linearity. The entire computation pipeline of LASK can be summarized as:

$$Y = \phi(\mathcal{B}(\mathcal{F}_{\text{PW}}^{1\times1}(\mathcal{L}(\sum_{i=1}^{3} S_i \odot \mathcal{F}_{\text{DW}}^{k_i}(X))))) \tag{11}$$

where $\phi(\cdot)$ denotes the SiLU function, $\mathcal{B}$ represents batch normalization, $\mathcal{F}_{\text{PW}}^{1\times1}(\cdot)$ represents 1×1 point-wise convolution.

The output Y maintains the same spatial resolution and channel configuration as the standard convolutional layer, allowing LASK to seamless integration into existing YOLO architectures. In this work, the proposed LASK module is selectively integrated into the P3 and P4 stages of the YOLOv12 backbone. Because P5 primarily captures high-level semantic information with low spatial resolution, it contributes little to fine-grained localization, whereas P1 and P2 mainly contain noisy low-level textures with insufficient representational depth; therefore, applying LASK in these stages yields limited improvement for small-object feature extraction and is computationally inefficient.

3.4. Dual-Stream Attention Bridge

To enhance small-object detection on UAV platforms with limited computational budgets, the neck design in this work follows a progressive strategy combining feature path simplification strategy and the proposed Dual-Stream Attention Bridge (DSAB). First, as shown in Figure 1, the feature path simplification strategy retains a pure top-down feature fusion pathway, replacing the bidirectional aggregation in PANet to prevent the over-propagation and dilution of fine-grained details during feature transmission while significantly reducing computational overhead. Then, the proposed DSAB injects high-level semantic guidance from P5 into the low-level feature map P3. Through semantic-guided foreground and background routing and lightweight multi-scale context aggregation, the DSAB selectively enhances informative regions and suppresses irrelevant background noise, enabling the network to focus on fine structural boundaries and small-object regions. In this way, the feature path simplification strategy guarantees efficient and undistorted information flow, while DSAB complements it by reintroducing rich semantics in a targeted and controllable manner. Together, these two components form a balanced neck architecture that maintains computational efficiency and substantially improves detection accuracy for small objects.

As shown in Figure **5**, given the high-level semantic feature map P_5 and the low-level feature map P_3, our DSAB aims to generate an enhanced representation E_3. We first align the high-level feature and generate a semantic attention map via sigmoid activation. Then the attention map is then used to derive the foreground stream A^{F} and background stream A^{B}:

 doi:10.20944/preprints202511.0375.v1

$$\begin{cases} \mathrm{A}^{\mathrm{F}} = \sigma(\mathcal{U}(\mathrm{Conv}(\mathrm{P}_5))) \\ \mathrm{A}^{\mathrm{B}} = \boldsymbol{E} - \mathrm{A}^{\mathrm{F}} \end{cases} \tag{12}$$

where $\mathrm{Conv}(\cdot)$ represents the convolution operation, $\mathcal{U}(\cdot)$ denotes the upsample operation and σ is the sigmoid function producing a spatial attention prior.

Low-level features P_3 are modulated into two semantic streams. Then each stream is processed through a Lite-ASPP module—comprising three depthwise separable convolutions with different dilation rates—to extract multi-scale contextual information., which is crucial for small-object recognition:

$$\mathrm{E}_3 = \Phi([\bigcup_{i=1}^{3} D_i\,(\mathrm{P}_3 \odot \mathrm{A}^{\mathrm{F}}), \bigcup_{i=1}^{3} D_i\,(\mathrm{P}_3 \odot \mathrm{A}^{B})]) \tag{13}$$

where $\odot$ denotes elements-wise multiplication, $\bigcup_{i=1}^{3} D_i\,(\cdot)$ represents concatenation the all results of $\{D_i\,(\cdot)\}_{i=1}^{3}$ operation. $\{D_i\,(\cdot)\}_{i=1}^{3}$ denotes a depthwise separable convolution with dilation rate of 1, 3, and 5. And They all have 3×3 kernels with 128 channels followed by a SiLU layer. The operator $[\cdot,\cdot]$ denotes channel-wise concatenation. $\Phi(\cdot)$ denotes the final fusion operator , where the concatenation feature are passed through a 3×3 convolution followed by a SiLU activation.

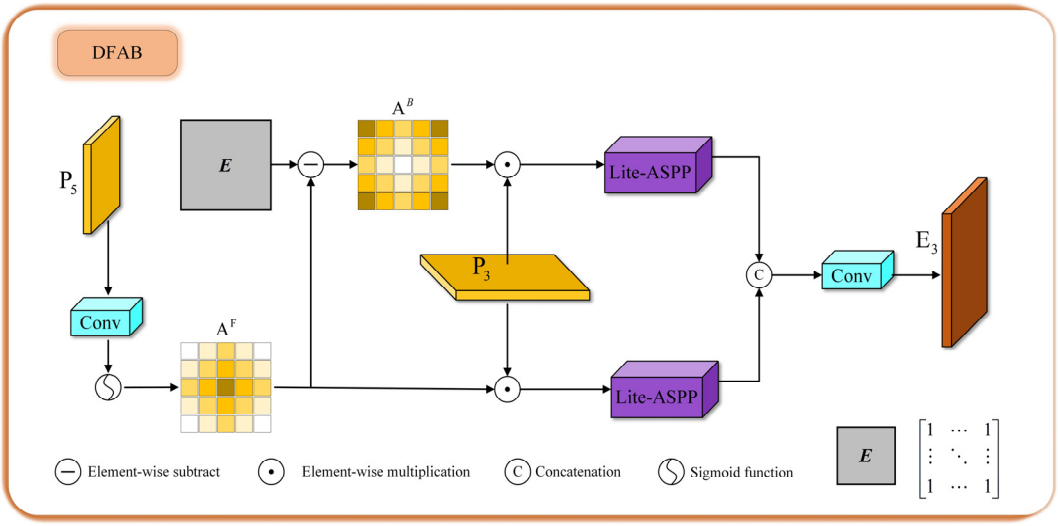

Figure 5. The object concealment failure results.

4. Experiment

4.1. Settings

- **Hardware/Software Environment.** The following experiments were performed on a computer with NVIDIA RTX 3090 GPU, 24 GB of memory, and Python code running on Ubuntu 24.04. With CUDA 11.6, PyTorch 1.12 and other commonly used deep learning and image processing libraries;
- **Dataset.** The VEDAI [46] dataset contains a total of 1,246 pairs of RGB and IR images, covering nine object categories such as cars, trucks, vans, tractors, pick-ups, and other road vehicles. The images were captured by UAV and aircraft platforms over semi-urban and rural areas, featuring complex backgrounds such as roads, buildings, vegetation, and shadowed regions. Each image has a resolution of 1024×1024 pixels, and the objects of interest are generally small in scale, often occupying less than 1% of the total image area. These characteristics make the dataset particularly challenging for small-object detection, as the vehicles exhibit significant variations in size, shape, and orientation, and are often affected by background clutter, illumination variations, and low inter-class contrast;

 doi:10.20944/preprints202511.0375.v1

- **Implementation Details.** We trained the network using the stochastic gradient descent (SGD) optimizer with the following hyperparameters. The initial learning rate is set to 0.01, the weight decay to 0.0005, and the momentum to 0.937. The model is trained for 300 iterations with a batch size of 8, and the input image size is set to 1024;
- **Evaluation Metrics.** In this work, we use Precision, Recall, and mean Average Precision (mAP) and Parameters as evaluation metrics to assess the performance of the detector as defined below:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{14}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{15}$$

$$\text{AP}_c = \int_0^1 P(r)dr \tag{16}$$

$$\text{mAP} = \frac{1}{N}\sum_{c=1}^{N} \text{AP}_c \tag{17}$$

where TP, FP, and FN denote true positive, false positive, and false negative, respectively. AP_c represents average precision for class c. $P(r)$ is the precision at the recall r and N is the number of classes.

4.2. Ablation Study

To assess the performance of the proposed method, we conducted ablation experiments on the VEDAI dataset, comparing multiple models across various metrics and marking the best results in bold.

4.2.1. Comparisons of Different Fusion Stages and Approaches

To evaluate the effectiveness of the proposed fusion method, we compared feature concatenation (Concat) with our TAFNet across six different detectors. Additionally, the various intermediate fusion strategies and the late fusion strategy were compared, with the network architecture illustrated in Figure 6. The experimental results of different fusion methods are presented in Table 1.

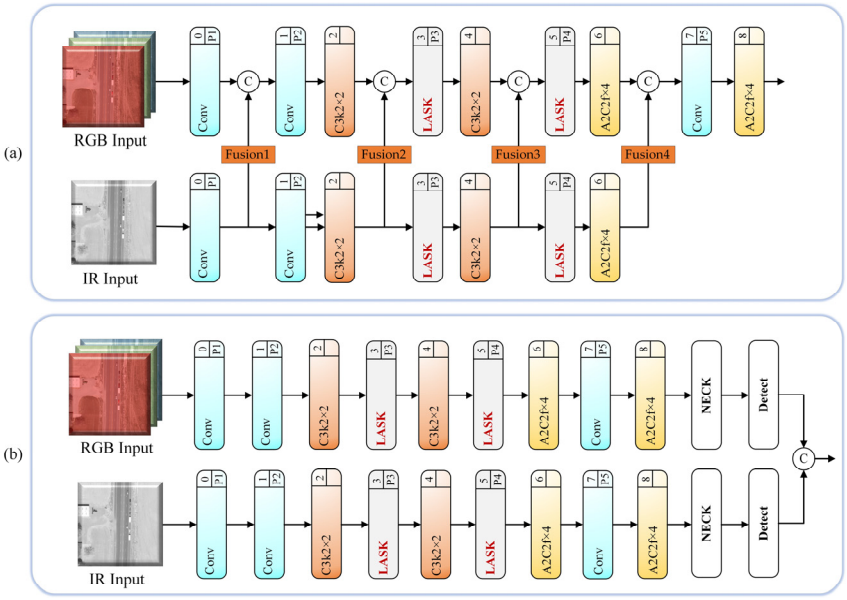

Figure 6. The network architecture of different fusion strategies. (a) Intermediate fusion strategies. (b) Late fusion strategy.

 doi:10.20944/preprints202511.0375.v1

Model size and inference performance were assessed using parameter counts (Params) and GFLOPs. Experimental results indicate that, compared with the early fusion method based on simple feature concatenation, our proposed approach achieves a noticeable improvement in detection accuracy with only a marginal increase in the number of parameters and GFLOPs. By evaluating multiple YOLO-based baseline models, we observe that the proposed TAFNet on the vast majority of baseline models consistently outperforms the simple concatenation-based early fusion strategy in both mAP_{50} and $mAP_{50:95}$ metrics, achieving the best detection performance. Specifically, Ours + TAFNet improves precision, recall, mA_{P50}, and $mAP_{50:95}$ by 4.5%, 4.2%, 3.1%, and 2.1%, respectively, compared with Our + Concat, indicating a significant performance improvement.

Moreover, our experimental results reveal a clear performance gap between the two early fusion strategies and the intermediate and late fusion strategies in terms of both detection accuracy and computational efficiency. These findings indicate that early fusion strategies are more effective in preserving low-level features, such as edges and textures, which are crucial for small object detection in UAV aerial imagery. In contrast, intermediate fusion tends to introduce redundant information and feature loss, potentially obscuring the representation of small objects. Late fusion fails to sufficiently exploit the complementary information between modalities while incurring additional computational overhead. Overall, these results demonstrate that the proposed TAFNet early fusion strategy achieves an excellent trade-off between detection accuracy and computational cost, making it particularly suitable for multimodal UAV-based small object detection tasks

Table 1. Comparison results of different fusion stages and approaches on the VEDIA datasets.

Method		Pre	Rec	mAP_{50}	$mAP_{50:95}$	Params	GFLOPs
Early Fusion	YOLOv5n+Concat	60.6	58.2	63.6	37.2	2.189M	6.0
	YOLOv5n+TAFNet	62.9	56.0	63.4	39.9	2.307M	6.4
	YOLOv6n+Concat	57.0	56.3	56.9	35.9	4.160M	11.7
	YOLOv6n+TAFNet	59.1	61.0	60.6	39.1	4.235M	11.9
	YOLOv8n+Concat	65.4	52.3	60.9	38.5	2.691M	7.0
	YOLOv8n+TAFNet	66.0	54.8	61.9	40.6	3.007M	7.4
	YOLOv10n+Concat	68.6	50.4	60.2	38.3	2.710M	8.3
	YOLOv10n+TAFNet	**71.6**	52.3	62.0	39.3	2.698M	8.7
	YOLOv11n+Concat	62.0	59.6	61.6	40.6	2.412M	6.0
	YOLOv11n+TAFNet	62.6	60.4	62.8	41.2	2.728M	6.4
	YOLOv12n+Concat	54.6	59.3	60.2	37.3	2.509M	5.9
	YOLOv12n+TAFNet	57.5	62.6	62.4	41.2	2.810M	6.3
	Ours+Concat	64.3	59.5	64.1	42.2	4.231M	8.0
	Ours+TAFNet	68.8	**63.7**	**67.2**	**44.3**	4.518M	8.4
Intermediate Fusion	Fusion1	60.2	59.3	62.1	39.8	4.639M	39.0
	Fusion2	62.6	59.6	62.7	40.7	4.654M	13.0
	Fusion3	61.7	61.4	63.1	41.2	4.827M	13.8
	Fusion4	67.3	60.8	63.6	41.5	5.297M	14.6
Late Fusion		60.2	54.3	59.8	37.2	11.709M	16.8

Note:Bold values indicate the best results.

4.2.2. Ablation Study of Each Component

Table 2 presents a comprehensive ablation study evaluating the individual and combined contributions of the proposed LASK, DSAB, and the effect of removing the PANet on the TAF-YOLO model's performance. The metrics considered include model Params, GFLOPs, Precision (Pre), Recall (Rec), mAP@0.5 (mAP_{50}), and mAP@0.5:0.95 ($mAP_{50:95}$).

As shown in Table 2, Adding LASK to the baseline leads to a notable improvement in Precision (from 57.5% to 61.8%). While mAP_{50} and $mAP_{50:95}$ show a slight increase (0.4% and 0.2% respectively), surprisingly, the GFLOPs slightly reduce from 6.3 to 6.1, despite a minor increase in parameters (from

 doi:10.20944/preprints202511.0375.v1

2.810M to 3.024M). This suggests LASK enhances the model's ability to dynamically select informative receptive fields while maintaining edge and texture integrity. Removing PANet from the baseline significantly reduces model complexity, decreasing parameters from 2.810M to 1.831M (a 34.8% reduction) and GFLOPs from 6.3 to 4.9 (a 22.3% reduction). Crucially, removing PANet also leads to a slight improvement in mAP_{50} (from 62.4% to 63.3%) and $mAP_{50:95}$ (from 41.2% to 41.9%). This indicates that the original PANet architecture is inefficient and introduces redundant features that are not suitable for small-object detection tasks. Removing it can improve efficiency and slightly enhance performance. When PANet is removed, adding LASK further improves Precision (from 58.8% to 62.1%), and also yields a marginal increase in mAP_{50} (0.4%) and $mAP_{50:95}$ (0.5%). The GFLOPs again show a slight reduction (from 4.9 to 4.7), reinforcing LASK's efficiency benefits. Adding DSAB while removing PANet significantly boosts mAP_{50} (from 62.4% to 66.9%) and $mAP_{50:95}$ (from 41.2% to 43.6%). This indicates that DSAB can effectively utilize high-level semantics to guide low-level features enabling the network to better perceive fine structural boundaries and small object regions. However, this requires acceptable computational increase : parameters increase to 4.305M and GFLOPs to 8.6. The best performing configuration, which incorporates both LASK and DSAB while removing PANet, achieves the highest scores across all accuracy metrics: Precision (64.8%), Recall (62.7%), mAP_{50} (67.2%), and $mAP_{50:95}$ (44.3%).

Table 2. The ablation experiment TAF-YOLO.

	LASK	DSAB	Remove PANet	Params	GFLOPs	Pre	Rec	mAP_{50}	$mAP_{50:95}$
1				2.810M	6.3	57.5	62.6	62.4	41.2
2	√			3.024M	6.1	61.8	62.2	62.8	41.4
3		√	√	4.305M	8.6	62.7	62.5	66.9	43.6
4			√	1.831M	4.9	58.8	62.0	63.3	41.9
5	√		√	2.044M	4.7	62.1	60.8	63.7	42.4
6	√	√	√	4.518M	8.4	**64.8**	**62.7**	**67.2**	**44.3**

Note:Bold values indicate the best results.

4.3. Comparison Experiment With Previous Methods

We compared the performance of state-of-the-art methods, including YOLOv5 [28], YOLOv6 [47], YOLOv8 [29], YOLOv10 [30], YOLOv11 [31], YOLOv12 [32], CFT [16], DEYOLO [48], and our proposed TAF-YOLO method on the VEDAI dataset across various scenarios. Figure 7 presents some visual comparison, where the red circles indicate missed detection, the yellow circles represent false detection, and the blue circles highlight false alarm. It can be observed that under scenarios with complex backgrounds or occlusions, other methods tend to produce a considerable number of missed and false detections. In UAV aerial imagery, vehicles usually occupy small regions, and their appearances across different categories often exhibit high similarity. Although classes such as car, pickup, camping, and truck are relatively abundant in the dataset and achieve better detection precision overall, large variations in object scale and inter-class resemblance still make accurate classification challenging. Most existing algorithms can accurately locate vehicles, but they have difficulty distinguishing fine-grained differences, which leads to frequent misclassifications, especially between cars and pickups.

The visualization results clearly demonstrate the superior performance of our proposed TAF-YOLO. It effectively addresses the critical challenges of small object detection, significant scale variation, and high inter-class similarity in UAV aerial imagery. Unlike other methods that struggle with missed detections, frequent misclassifications (especially between car and pickup), and false alarms in complex or cluttered backgrounds, TAF-YOLO consistently delivers more accurate and robust detection outcomes. This indicates that our method's architecture and fusion strategy are particularly well-suited for extracting discriminative features necessary for precise localization and fine-grained classification in these demanding scenarios.

 doi:10.20944/preprints202511.0375.v1

Table 3 summarizes the performance of the above methods. The MF model achieved significantly better AP values than the single modal (RGB or IR) model in most categories, indicating that the proposed TAFNet can effectively utilize early feature fusion strategies. However, it is worth noting that for the "car" category, the detection performance of the fusion model is actually lower than using only RGB modes. This phenomenon may be related to the information differences between different modalities and the performance of specific categories (such as cars) under different lighting conditions. In some cases, infrared information may not provide sufficient details or have significant differences from visible light images, thereby affecting the detection accuracy after fusion. However, overall, the improvement in mAP_{50} and $mAP_{50:95}$ still indicates that TAFNet exhibits superior performance in most categories. Meanwhile, the increase in parameter count and computational complexity is very limited, indicating that the method has achieved an ideal balance between accuracy and efficiency.

Table 3. The quantitative comparison of different methods on VEDAI.

Method		Car	Pickup	Camping	Truck	Others	Tractor	Boat	Van	mAP_{50}	$mAP_{50:95}$	Params	GFLOPs
YOLOv5n	RGB	88.3	69.4	60.6	59.2	34.6	49.7	65.1	52.9	60.0	37.4	2.189M	5.9
	IR	85.7	64.9	67.5	60.3	22.2	53.6	35.5	65.7	56.9	34.5	2.189M	5.9
	Multi	88.1	72.0	65.2	65.4	47.2	53.8	63.3	51.8	63.4	39.9	2.307M	6.4
YOLOv6n	RGB	87.1	72.0	60.1	59.2	44.2	55.2	35.1	55.3	58.5	36.1	4.160M	11.6
	IR	86.7	65.6	52.8	65.6	22.2	55.1	14.5	52.2	51.8	31.7	4.160M	11.6
	Multi	84.5	68.7	60.9	59.4	53.7	60.0	27.8	67.9	60.6	39.1	4.335M	11.9
YOLOv8n	RGB	87.8	76.1	55.0	54.7	38.6	62.1	51.5	45.3	58.9	35.5	2.691M	6.9
	IR	85.2	66.6	63.6	63.3	30.5	41.4	35.7	52.5	54.8	33.4	2.691M	6.9
	Multi	83.1	73.6	61.8	64.3	41.6	68.1	44.9	39.0	61.9	40.6	3.007M	7.4
YOLOv10n	RGB	84.9	69.9	61.3	59.9	17.6	49.4	48.9	50.5	55.3	33.9	2.710M	8.2
	IR	81.5	68.5	55.1	53.8	23.8	36.9	30.9	51.9	50.3	31.2	2.710M	8.2
	Multi	81.2	66.9	61.6	63.8	37.9	48.5	57.6	78.4	62.0	39.3	2.842M	8.7
YOLOv11n	RGB	88.8	71.7	75.9	57.3	31.4	59.4	40.7	49.3	59.3	35.4	2.412M	5.9
	IR	84.7	69.1	64.9	73.6	30.4	37.3	39.8	47.7	55.9	33.9	2.412M	5.9
	Multi	84.9	74.1	67.3	72.0	44.8	72.1	62.2	48.8	62.8	41.2	2.728M	6.6
YOLOv12n	RGB	87.0	71.6	63.5	61.1	42.8	56.9	46.7	42.1	59.0	35.2	2.509M	5.8
	IR	80.8	64.7	57.0	66.0	22.4	34.6	29.1	37.1	49.0	29.9	2.509M	5.8
	Multi	83.3	73.4	75.7	58.0	45.6	60.4	54.2	59.9	62.4	41.2	2.810M	6.3
CFT	Multi	85.6	73.2	65.2	65.6	44.3	64.2	54.2	60.8	64.5	42.1	47.06M	117.2
DEYOLOn	Multi	85.2	70.5	70.0	68.3	50.3	67.0	60.2	61.5	65.7	43.2	6.09M	-
TAF-YOLO (Ours)	RGB	88.5	70.4	63.4	62.1	43.2	57.2	52.2	45.6	62.2	40.4	4.231M	7.9
	IR	84.5	68.7	60.9	59.4	30.2	39.6	27.8	39.6	60.3	38.3	4.231M	7.9
	Multi	87.0	72.3	75.7	71.3	47.6	65.4	67.9	62.9	67.2	44.3	4.518M	8.4

Note:Bold values indicate the best results, and "-" indicates the results is not reported.

The visualization results of several failed detections presented in Figure 8 reveal the limitations of TAF-YOLO in highly complex scenarios. Missed detections typically occur when objects are extremely small or blend seamlessly with the background. A representative example is the "tractors" in Figure 8(a), located at the top-left corner of the image along a dirt path, which remains undetected due to its minute size and similar texture to the surrounding ground. And as shown in Figure 8(d), a "car" is erroneously identified as a "pickup," clearly demonstrating the model's difficulty in fine-grained classification, particularly between visually similar vehicles like cars and pickups, where subtle structural differences are flattened and obscured from an overhead view. In summary, the failure cases indicate the extreme complexity and inherent challenges of object detection in UAV aerial imagery, encompassing the diminutive size of objects, the intrinsic high visual similarity

 doi:10.20944/preprints202511.0375.v1

among different vehicle classes, and the strong interference from diverse and complex backgrounds (e.g., farmland, urban structures, shadows, road textures).

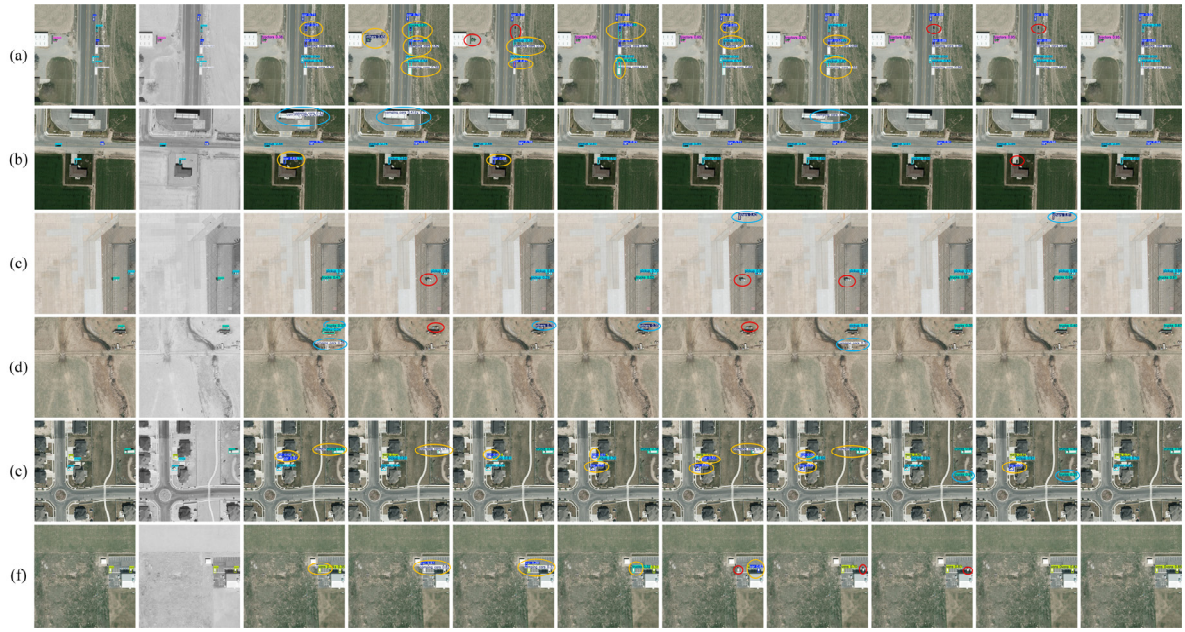

Figure 7. The visual comparison of different inpainting methods. Left to right: RGB Ground Truth, IR Ground Truth, YOLOv5, YOLOv6, YOLOv8, YOLOv10, YOLOv11, YOLOv12, CFT, DEYOLO, and Ours. The red circles indicate missed detection, the yellow circles represent false detection, and the blue circles highlight false alarm.

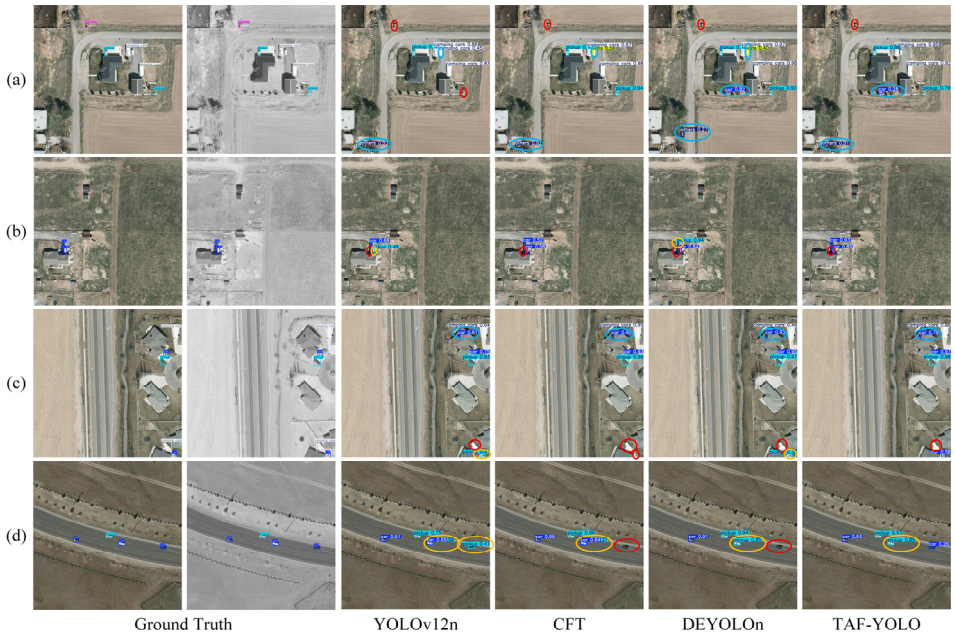

Figure 8. The failure cases of MF detection methods. The red circles indicate missed detection, the yellow circles represent false detection, and the blue circles highlight false alarm.

4.4. Generalization Experiment

To validate the generalization capability of the proposed TAF-YOLO, we conduct a comparative evaluation of TAF-YOLO and other detectors on the public DroneVehicle dataset [49]. DroneVehicle captures complex airborne imaging scenarios, where vehicles exhibit significant variations in size, orientation, and background clutter due to changes in drone altitude and camera tilt. The dataset provides paired RGB and IR images, enabling the exploration of MF detector's robustness under diverse environmental conditions. The dataset encompasses diverse illumination situations (daytime, dusk, and

 doi:10.20944/preprints202511.0375.v1

nighttime) and viewing angles, simulating real-world drone surveillance tasks. Such complementary characteristics make DroneVehicle well-suited for evaluating MF detecor and generalization ability across domains. For our experiments, we follow the common split setting and select 2800 image pairs for training, 800 for validation, and 400 for testing. The annotations cover five vehicle categories—car, truck, bus, van, and freight-car—each appearing at varying scales and densities. The presence of occlusions, dense traffic, and multi-scale objects poses additional challenges for detection algorithms.

Table 4 presents in detail the comparison results of different algorithms on the DroneVehicle dataset, while Figure 9 highlights on the visualization experimental results of CFT, DEYOLO, and TAF-YOLO in night scenes, ground texture interference scenes, and low light scenes. Experimental results on the public dataset DroneVehicle show that our proposed TAF-YOLO achieved mAP_{50} and $mAP_{50:95}$ of 65.8% and 46.7%, respectively. This val-idates the efficacy and superiority of the proposed approach for detecting small objects in UAV aerial images, further demonstrating its excellent generalization capability. The visualization results demonstrate that the proposed TAF-YOLO excels in detecting objects under challenging conditions, such as night scenes, ground texture interference, and low-light environments. Compared to the state-of-the-art methods, CFT and DEYOLO, TAF-YOLO significantly reduces false detections and false alarms, while also improving both positional accuracy and confidence levels. In addition, TAF-YOLO achieves a favorable balance between accuracy and computational efficiency. With only 4.518M parameters and 8.4 GFLOPs, it delivers performance comparable to larger models such as CFT and DEYOLO, demonstrating the proposed method's potential for deployment in real-time UAV detection tasks with limited computational resources.

Table 4. The quantitative comparison of different methods on DroneVehicle.

Method		Cars	Trucks	Buses	Vans	Freight cars	mAP_{50}	$mAP_{50:95}$	Params	GFLOPs
YOLOv5n	Multi	91.8	46.3	84.3	45.3	44.6	62.1	41.4	2.307M	6.4
YOLOv8n	Multi	92.0	50.4	84.3	46.6	45.2	62.9	42.1	3.007M	7.4
YOLOv10n	Multi	91.3	44.8	83.6	40.2	44.4	60.3	41.1	2.842M	8.9
YOLOv11n	Multi	91.3	36.4	80.6	33.7	41.8	57.4	38.3	2.728M	6.6
YOLOv12n	Multi	91.4	45.1	84.5	42.3	45.0	61.8	40.4	2.810M	6.3
CFT	Multi	**92.4**	55.5	88.6	44.6	**46.3**	65.4	45.7	47.06M	117.2
DEYOLOn	Multi	91.0	51.8	85.2	44.1	46.9	64.4	43.6	6.09M	-
TAF-YOLO	Multi	91.7	**58.6**	**89.2**	**46.9**	44.3	**65.8**	**46.7**	4.518M	8.4

Note:Bold values indicate the best results, and "-" indicates the results is not reported.

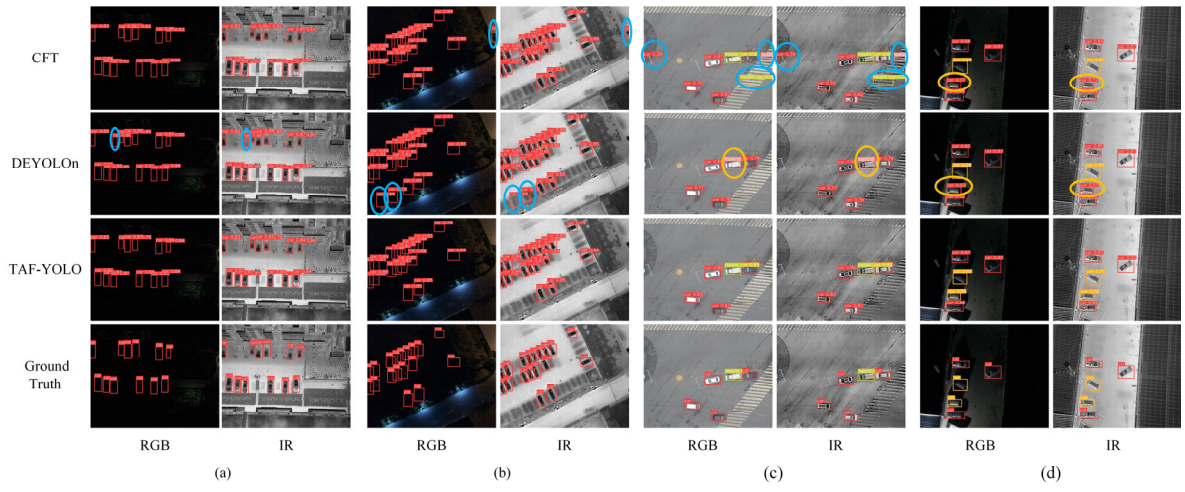

Figure 9. Generalization experiments in different scenarios on DroneVehicle dataset. (a) and (b) are night scenes, (c) is a ground texture interference scene, and (d) is a low light scene. The red circles indicate missed detection, the yellow circles represent false detection, and the blue circles highlight false alarm.

 doi:10.20944/preprints202511.0375.v1

5. Discussion

The proposed TAF-YOLO effectively leverages TAFNet for pixel-level and channel-level adaptive fusion of visible and infrared inputs, enabling comprehensive integration of complementary information while minimizing redundancy. It demonstrates strong detection performance in nighttime, background-interference, and low-light environments, reducing both miss detection and false alarm rates. However, its performance in extreme scenarios still requires further improvement. As shown in Figure 10 (a) and (b), in extreme modality imbalance scenarios, the visible light modality fails to provide effective information, while the infrared modality exhibits strong background noise, leading to false alarms. In the multimodal information scarcity scenarios shown in Figure 10 (c) and (d), both modalities fail to provide useful information, resulting in missed detections.

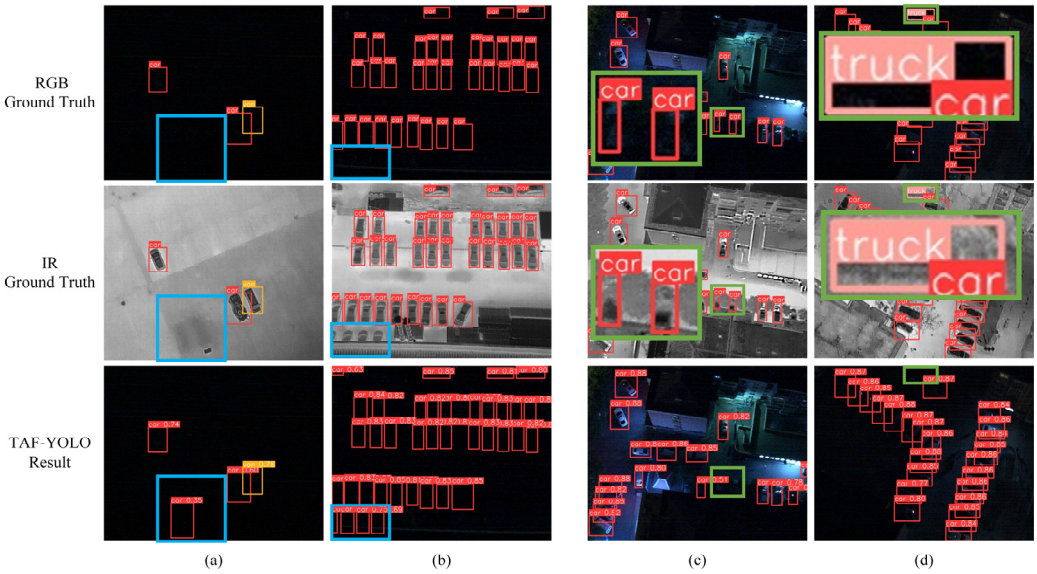

Figure 10. Multimodal fusion under extreme scenarios. (a) and (b) depict a modality imbalance scenario. (c) and (d) illustrate a multimodal information scarcity scenario.

And the LASK module dynamically adjusts receptive fields through multi-scale depthwise convolutions and spatial attention, preserving fine edge and texture details during downsampling. By replacing the traditional convolution operations in P3 and P4 of the original baseline model with this approach, we achieve promising results. However, we have not yet established that replacing the baseline convolutional blocks with the LASK module will consistently outperform the original convolutions across different datasets and architectures. Additional experiments are needed to verify LASK's robustness and generalization as a direct replacement for standard convolutional layers.

6. Conclusions

In this paper, we presented TAF-YOLO, an efficient multimodal small-object detection framework tailored for UAV aerial imagery. By performing early adaptive fusion of visible and infrared modalities through the proposed TAFNet, the framework effectively integrates complementary texture and thermal information while suppressing redundancy. The LASK module dynamically enhances receptive field adaptability and preserves fine-grained edge and texture features, significantly improving the perception of small objects. Furthermore, the incorporation of DSAB and a simplified feature fusion path strengthens semantic interaction and localization precision while maintaining lightweight computational costs. Extensive experiments on the VEDAI dataset demonstrate that TAF-YOLO achieves superior performance in both detection accuracy and efficiency, confirming its suitability for real-time, all-weather UAV perception. However, in extreme detection scenarios characterized by strong background interference and extremely small inter-class differences, the algorithm still requires enhanced feature extraction and background suppression capabilities. Future research will focus on further enhancing the

 doi:10.20944/preprints202511.0375.v1

performance of TAF-YOLO in extreme scenarios with significant modal imbalance to improve generalization across diverse UAV platforms and complex environments, extending the applicability of TAF-YOLO to broader multimodal UAV remote sensing tasks.

Author Contributions: Conceptualization, R.L. and Y.Y.; methodology, Z.Z.; software, S.W.; validation, R.L., Z.Z. and S.W.; investigation, R.L.; resources, Z.Z.; writing—original draft preparation, Z.Z.; writing—review and editing, Z.Z and R.L.; visualization, J.Z.; supervision, R.L.; project administration, Z.Z.; funding acquisition, X.Y. and R.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the National Natural Science Foundation of China under grant number 62276274 and 62573423, in part by the Key R&D Program of Shaanxi Province 2024CY2-GJHX-42, and in part by the Shannxi Sanqin Elite Special Support Program 2024-SQ-001.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

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

Abbreviations

The following abbreviations are used in this manuscript:

UAV	Unmanned Aerial Vehicle
CNNs	Convolutional Neural Networks
IR	Infrared
MF	Multimodal Fusion
VIF	Visible and Infrared Image Fusion
NMS	Non-Maximum Suppression
TAFNet	Two-branch Adaptive Fusion Network
LASK	Large Adaptive Selective Kernel
DSAB	Dual-Stream Attention Bridge

References

1. Osco, L.P.; Marcato Junior, J.; Marques Ramos, A.P.; et al. A review on deep learning in UAV remote sensing. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *102*, 102456, doi:10.1016/j.jag.2021.102456.
2. Aasen, H.; Honkavaara, E.; Lucieer, A.; et al. Quantitative Remote Sensing at Ultra-High Resolution with UAV Spectroscopy: A Review of Sensor Technology, Measurement Procedures, and Data Correction Workflows. *Remote Sens.* **2018**, *10*, doi:10.3390/rs10071091.
3. Bo, W.; Liu, J.; Fan, X.; et al. BASNet: Burned Area Segmentation Network for Real-Time Detection of Damage Maps in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1-13, doi:10.1109/TGRS.2022.3197647.
4. Yu, X.; Jiang, T.; Zhu, Y.; et al. FEL-YoloV8: A New Algorithm for Accurate Monitoring Soybean Seedling Emergence Rates and Growth Uniformity. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1-17, doi:10.1109/TGRS.2025.3578800.
5. Xiao, Y.; Wang, J.; Zhao, Z.; et al. UAV Video Vehicle Detection: Benchmark and Baseline. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1-14, doi:10.1109/TGRS.2025.3534524.
6. Luo, M.; Zhao, R.; Zhang, S.; et al. IM-CMDet: An Intramodal Enhancement and Cross-Modal Fusion Network for Small Object Detection in UAV Aerial Visible-Infrared Imagery. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1-16, doi:10.1109/TGRS.2025.3615481.
7. Du, B.; Huang, Y.; Chen, J.; et al. Adaptive Sparse Convolutional Networks with Global Context Enhancement for Faster Object Detection on Drone Images. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 17-24 June 2023; pp. 13435-13444.
8. Wang, K.; Fu, X.; Huang, Y.; et al. Generalized UAV Object Detection via Frequency Domain Disentanglement. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 17-24 June 2023; pp. 1064-1073.

 doi:10.20944/preprints202511.0375.v1

9. Zou, Z.; Hu, X.; Zhong, P. Active Object Detection for UAV Remote Sensing via Behavior Cloning and Enhanced Q-Network With Shallow Features. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1-16, doi:10.1109/TGRS.2025.3563486.
10. Jiao, Z.; Wang, M.; Qiao, S.; et al. Transformer-Based Object Detection in Low-Altitude Maritime UAV Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1-13, doi:10.1109/TGRS.2025.3594085.
11. Zhang, X.; Demiris, Y. Visible and Infrared Image Fusion Using Deep Learning. *IEEE Trans. Pattern Anal. Mach. Intelle.* **2023**, *45*, 10535-10554, doi:10.1109/TPAMI.2023.3261282.
12. Xu, S.; Chen, X.; Li, H.; et al. Airborne Small Target Detection Method Based on Multimodal and Adaptive Feature Fusion. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1-15, doi:10.1109/TGRS.2024.3443856.
13. Liu, J.; Fan, X.; Jiang, J.; et al. Learning a Deep Multi-Scale Feature Ensemble and an Edge-Attention Guidance for Image Fusion. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 105-119, doi:10.1109/TCSVT.2021.3056725.
14. Li, Q.; Zhang, C.; Hu, Q.; et al. Confidence-Aware Fusion Using Dempster-Shafer Theory for Multispectral Pedestrian Detection. *IEEE Trans. Multimedia.* **2023**, *25*, 3420-3431, doi:10.1109/TMM.2022.3160589.
15. Park, S.; Vien, A.G.; Lee, C. Cross-Modal Transformers for Infrared and Visible Image Fusion. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 770-785, doi:10.1109/TCSVT.2023.3289170.
16. Qingyun, F.; Dapeng, H.; Zhaokui, W.J.a.p.a. Cross-modality fusion transformer for multispectral object detection. **2021**, *arXiv:2111.00273*.
17. Girshick, R.; Donahue, J.; Darrell, T.; et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, 23-28 June 2014; pp. 580-587.
18. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), 7-13 Dec. 2015; pp. 1440-1448.
19. Ren, S.; He, K.; Girshick, R.; et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137-1149, doi:10.1109/TPAMI.2016.2577031.
20. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 18-23 June 2018; pp. 6154-6162.
21. Liu, W.; Anguelov, D.; Erhan, D.; et al. SSD: Single Shot MultiBox Detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14; pp. 21-37.
22. Lin, T.Y.; Goyal, P.; Girshick, R.; et al. Focal Loss for Dense Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), 22-29 Oct. 2017; pp. 2999-3007.
23. Carion, N.; Massa, F.; Synnaeve, G.; et al. End-to-End Object Detection with Transformers. In Proceedings of the Computer Vision – ECCV 2020, Cham, 2020//; pp. 213-229.
24. Zhu, X.; Su, W.; Lu, L.; et al. Deformable DETR: Deformable transformers for end-to-end object detection. **2020**, *arXiv:2010.04159*.
25. Liu, Z.; Lin, Y.; Cao, Y.; et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 10-17 Oct. 2021; pp. 9992-10002.
26. Dosovitskiy, A.J.a.p.a. An image is worth 16x16 words: Transformers for image recognition at scale. **2020**, *arXiv:2010.11929*.
27. Redmon, J.; Divvala, S.; Girshick, R.; et al. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 27-30 June 2016; pp. 779-788.
28. Tian, Z.; Shen, C.; Chen, H.; et al. FCOS: A Simple and Strong Anchor-Free Object Detector. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 1922-1933, doi:10.1109/TPAMI.2020.3032166.
29. Varghese, R.; S, M. YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness. In Proceedings of the 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), 18-19 April 2024; pp. 1-6.
30. Wang, A.; Chen, H.; Liu, L.; et al. Yolov10: Real-time end-to-end object detection. *2024 Advances in neural information processing systems (NeurIPS)* **2024**, *37*, 107984-108011.

31. Khanam, R.; Hussain, M.J.a.p.a. Yolov11: An overview of the key architectural enhancements. **2024**, *arxiv:2410.17725*.
32. Tian, Y.; Ye, Q.; Doermann, D.J.a.p.a. Yolov12: Attention-centric real-time object detectors. **2025**, *arXiv:2502.12524*.
33. Wu, S.; Lu, X.; Guo, C.; et al. MV-YOLO: An Efficient Small Object Detection Framework Based on Mamba. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1-14, doi:10.1109/TGRS.2025.3584955.
34. Liu, Y.; Cheng, X.; Xu, N.; et al. MFAE-YOLO: Multifeature Attention-Enhanced Network for Remote Sensing Images Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1-14, doi:10.1109/TGRS.2025.3583467.
35. Cao, J.; Bao, W.; Shang, H.; et al. GCL-YOLO: A GhostConv-Based Lightweight YOLO Network for UAV Small Object Detection. *Remote Sens.* **2023**, *15*, doi:10.3390/rs15204932.
36. Tang, Y.; Xu, T.; Qin, H.; et al. IRSTD-YOLO: An Improved YOLO Framework for Infrared Small Target Detection. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 1-5, doi:10.1109/LGRS.2025.3562096.
37. Hao, X.; Luo, S.; Chen, M.; et al. Infrared small target detection with super-resolution and YOLO. *Opt. Laser Technol.* **2024**, *177*, 111221, doi:10.1016/j.optlastec.2024.111221.
38. Caesar, H.; Bankiti, V.; Lang, A.H.; et al. nuScenes: A Multimodal Dataset for Autonomous Driving. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 13-19 June 2020; pp. 11618-11628.
39. Hong, D.; Gao, L.; Yokoya, N.; et al. More Diverse Means Better: Multimodal Deep Learning Meets Remote-Sensing Imagery Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 4340-4354, doi:10.1109/TGRS.2020.3016820.
40. Wang, Z.; Liao, X.; Yuan, J.; et al. CDC-YOLOFusion: Leveraging Cross-Scale Dynamic Convolution Fusion for Visible-Infrared Object Detection. *IEEE Trans. Intell. Veh.* **2025**, *10*, 2080-2093, doi:10.1109/TIV.2024.3443264.
41. Shen, J.; Chen, Y.; Liu, Y.; et al. ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognit.* **2024**, *145*, 109913, doi:10.1016/j.patcog.2023.109913.
42. Zeng, Y.; Liang, T.; Jin, Y.; et al. MMI-Det: Exploring Multi-Modal Integration for Visible and Infrared Object Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 11198-11213, doi:10.1109/TCSVT.2024.3418965.
43. Tang, K.; Ma, Y.; Miao, D.; et al. Decision Fusion Networks for Image Classification. *IEEE Trans. Neural Netw. Learn, Syst.* **2025**, *36*, 3890-3903, doi:10.1109/TNNLS.2022.3196129.
44. Zhang, J.; Lei, J.; Xie, W.; et al. SuperYOLO: Super Resolution Assisted Object Detection in Multimodal Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1-15, doi:10.1109/TGRS.2023.3258666.
45. Chen, Y.; Da, F. YOLO-FMEN: Pixel-Level Image Fusion-Based Nighttime Object Detection Network. In Proceedings of the 2025 44th Chinese Control Conference (CCC), 28-30 July 2025; pp. 9179-9186.
46. Razakarivony, S.; Jurie, F. Vehicle detection in aerial imagery : A small target detection benchmark. *J. Vis. Commun. Image Represent.* **2016**, *34*, 187-203, doi:10.1016/j.jvcir.2015.11.002.
47. Li, C.; Li, L.; Jiang, H.; et al. YOLOv6: A single-stage object detection framework for industrial applications. **2022**, *arXiv:2209.02976*.
48. Chen, Y.; Wang, B.; Guo, X.; et al. DEYOLO: Dual-Feature-Enhancement YOLO for Cross-Modality Object Detection. In Proceedings of the Pattern Recognition, Cham, 2025//; pp. 236-252.
49. Sun, Y.; Cao, B.; Zhu, P.; et al. Drone-Based RGB-Infrared Cross-Modality Vehicle Detection Via Uncertainty-Aware Learning. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6700-6713, doi:10.1109/TCSVT.2022.3168279.