Essay

IV-YOLO: A Lightweight Dual-Branch Object Detection Network

Xin Yan , Dan Tian * , Dong Zhou , Chen Wang , Wenshuai Zhang

Posted Date: 28 August 2024

doi: 10.20944/preprints202408.2054.v1

Keywords: Dual-branch image object detection; IV-YOLO; Bi-directional pyramid feature fusion; Attention mechanism; Small target detection

 doi:10.20944/preprints202408.2054.v1

Essay

IV-YOLO: A Lightweight Dual-Branch Object Detection Network

Xin Yan, Dan Tian *, Dong Zhou , Cheng Wang and Wenshuai Zhang

University of Electronic Science and Technology of China, Chengdu 611731, China
* Correspondence: tiandan@uestc.edu.cn

Abstract: With the increasing demand for security surveillance, assisted driving, and remote sensing, target detection networks with rich environmental perception capabilities and high detection accuracy have become a research hotspot. However, detection techniques based on single-modality images face limitations in environmental adaptability, as they are easily affected by lighting conditions, as well as obstacles such as smoke, rain, and vegetation, leading to information loss and failing to meet the required accuracy. To address these issues, we propose a target detection network—IV-YOLO, which integrates features from both visible and infrared images. We built a dual-branch fusion network based on YOLOv8 (You Only Look Once v8) that combines infrared and visible images for target detection. On this foundation, we designed a Bidirectional Pyramid Feature Fusion structure (Bi-Fusion) to fully integrate complementary features from different modalities, reducing detection errors caused by feature redundancy and extracting fine-grained features of small targets through dual-modal fusion. Additionally, we developed the Shuffle-SPP structure, which combines channel and spatial attention mechanisms to enhance focus on deep features and further aggregate upsampling, extracting richer features. To improve the model's expressive power, we designed a loss function tailored for multi-scale target detection boxes, accelerating the network's convergence during training. Experimental results show that the proposed IV-YOLO network achieves an average precision (mAP) of 74.6% on the UAV remote sensing dataset, 75.4% on the KAIST pedestrian dataset, and 84.8% on the FLIR dataset. While ensuring detection accuracy, our model significantly reduces computational load, making it well-suited to meet real-time requirements. This architecture is expected to find broad application in fields such as autonomous driving, public safety, and others.

Keywords: dual-branch image object detection; IV-YOLO; bi-directional pyramid feature fusion; attention mechanism; small target detection

1. Introduction

Bimodal image detection has emerged as a significant direction in the field of computer vision, enhancing the performance and robustness of target detection systems by leveraging information from two different sensing modalities. Traditional single-modal image detection methods rely solely on information from one modality, making them susceptible to variations in lighting, occlusion, and environmental complexity, which can lead to decreased detection accuracy and increased error rates [1–3]. Bimodal image detection, however, can overcome the limitations of single-modal approaches, improving system performance across various complex scenarios.

As we all know,infrared images provide stable thermal signals in low-light, nighttime, or adverse weather conditions [4], while visible light images offer rich details and color information under normal lighting conditions [5]. Fusing information from these two modalities can significantly enhance the robustness of target detection. However, effectively merging infrared and visible light images is a complex task, requiring the establishment of meaningful associations and complementary relationships between the two modalities. The feature extraction and fusion processes for different modalities must overcome modal disparities and leverage the unique advantages of each modality, making the complementary nature of multimodal features particularly critical.

By exploring the complementary information between different sensing modalities, models can enhance the richness and diversity of target features, enabling more accurate identification and classification of various target types. This not only improves detection accuracy but also effectively

 doi:10.20944/preprints202408.2054.v1

reduces false positives and missed detections, resulting in superior performance in complex and dynamic real-world scenarios. Additionally, it involves designing appropriate network architectures and loss functions to optimize the model's learning and generalization capabilities.

In recent years, significant progress has been made in the field of object detection for dual-modal images. For instance, FusionNet [6] utilizes convolutional neural networks (CNNs) as the primary feature extractors and inputs the fused feature maps into an object detection network for further analysis. Dual-YOLO [7] introduces information fusion modules and fusion shuffle modules, enabling the network to use visible light features to complement infrared images, thereby enhancing detection accuracy and robustness. Guan et al. [8] employed deep convolutional neural networks to extract features from infrared and visible light images separately and then fused these features using a convolutional fusion module. Meanwhile, Kim et al. [9] improved overall detection performance through joint training methods.

However, existing methods still have four main limitations. First, precise control and interpretation of dual-modal feature fusion remain challenging, often resulting in redundant or insufficient information in the fused features. Second, during the forward propagation in fusion networks, deep-layer information is prone to loss, which can significantly affect detection accuracy, especially in complex scenarios [10,11]. Third, many datasets include targets of various scales, and extracting fine-grained features and achieving effective detection become extremely difficult under complex backgrounds or large object occlusions. Lastly, fusion networks typically involve a large number of parameters, leading to substantial model size and high computational costs, which in turn impacts the real-time performance of object detection.

To address the aforementioned issues, we propose a real-time object detection method based on a dual-branch fusion of visible light and infrared images. This method effectively overcomes the problems of poor environmental adaptability and information loss inherent in single-modal image detection, and tackles challenges such as feature disparity, information redundancy, and model optimization during the feature fusion process, achieving both efficient and accurate object detection. Experimental results demonstrate that the dual-branch image detection approach significantly improves the detection accuracy of multi-scale targets in complex environments while maintaining a minimal parameter count. Our contributions can be summarized as follows:

(1)Building on the highly efficient YOLOv8 [12] network, which excels in real-time object detection, we have developed a dual-branch network. This network includes one branch for extracting target features from infrared images and another branch for extracting target features from visible light images, and is named IV-YOLO. This approach effectively addresses the issues of complex backgrounds and target feature loss in single-modal image detection, and is capable of extracting fine-grained features of small targets from the network. While maintaining a low parameter count and high frame rate, it significantly enhances the detection accuracy of multi-scale targets in complex backgrounds.

(2)We proposed a method that integrates the Bi-Fusion module, using a bi-directional pyramid structure to fuse features extracted from infrared and visible light images at different levels in a weighted manner. This method effectively reduces redundant features, achieving optimal fusion of the features from both modalities.

(3)We proposed the Shuffle-SPP structure, which enhances the model's ability to detect multi-scale objects by extracting features from different scales through Spatial Pyramid Pooling (SPP). Subsequently, an efficient hierarchical aggregation mechanism is employed to fuse features from different levels, further improving feature representation. Based on this, features are divided into different parts, and by focusing on the details and positions of each part, the accuracy and precision of object detection are enhanced. This multi-level feature fusion, shuffling, and pooling operation effectively reduces information loss, retaining more critical information and thereby improving the overall performance of the network. Additionally, we introduced a loss function to enhance the network's attention to small objects and accelerate network convergence.

 doi:10.20944/preprints202408.2054.v1

(4)Our method achieved state-of-the-art results on the challenging KAIST [13] multispectral pedestrian dataset and the Drone Vehicle [14] dataset. Furthermore, experiments on the FLIR [15] multispectral object detection dataset also demonstrated the effectiveness and generality of the algorithm.

The remainder of this paper is structured as follows: Section 2 describes related work pertinent to our network. In Section 3, we detail the network architecture and methodology. Section 4 presents our experimental details and results, comparing them with state-of-the-art networks to validate the effectiveness of our approach. In Section 5, we summarize the research content and experimental findings.

2. Related Work

This section briefly reviews the classical methods of deep learning (DL) in real-time object detection, summarizes the research progress in multimodal image fusion techniques, and explores the structures of attention mechanisms.

2.1. Real-Time Object Detector

With the rapid advancement of deep learning, particularly convolutional neural networks (CNNs), real-time object detection technology has made significant strides in classifying and localizing objects under low-latency conditions. Modern real-time object detection methods primarily rely on deep learning models such as the YOLO (You Only Look Once) series, SSD [16] (Single Shot MultiBox Detector), and Fast R-CNN [17]. The YOLO series achieves object localization and classification through a single forward pass, while SSD directly predicts the classes and locations of multiple objects using a convolutional neural network, and Fast R-CNN detects objects by first generating candidate regions and then classifying and precisely localizing them.

Among these, the YOLO series has gradually become mainstream due to its efficient design. Versions like YOLOv1 [18], YOLOv2 [19], and YOLOv3 [20] have defined the classic detection architecture, which typically consists of three parts: the backbone, the neck, and the head. YOLOv4 [21] and YOLOv5 [22] introduced the CSPNet design to replace DarkNet, combined with data augmentation strategies, enhanced Path Aggregation Network (PAN), and more diversified model scales. YOLOv6 [23] introduced BiC and SimCSPSPPF in the neck and backbone, and adopted anchor-assisted training and self-distillation strategies. YOLOv7 [24] introduced the E-ELAN layer, while YOLOv8 proposed the C2f block for efficient feature extraction and fusion. YOLOv9 [25] further improved the architecture with the introduction of GELAN and incorporated PGI to enhance the training process. The latest YOLOv10 [26] introduced a continuous dual assignment method with NMS-free (Non-Maximum Suppression) training, offering competitive performance and the advantage of low inference latency.

2.2. Deep Learning-Based Multimodal Image Fusion

In the field of deep learning-based multimodal image fusion and detection, extensive research has been conducted to explore this issue in depth. Some studies use Convolutional Neural Networks (CNNs) to extract features from different modalities and perform detection through feature-level fusion methods. For example, Farahnakian, et al. [27] explored how to effectively fuse multimodal features using CNNs to improve detection accuracy and robustness. Additionally, Generative Adversarial Networks (GANs) have been widely applied in multimodal image fusion and detection to enhance model robustness and detection precision. Li,, et al.[28] proposed a multimodal GAN architecture that combines visible and thermal imaging, effectively fusing images from different sensors using GANs. Researchers have also explored methods that integrate multi-task learning with weak supervision and self-supervised learning to further optimize multimodal detection performance. For instance, FFD [29] proposed a self-supervised learning framework to accomplish fusion tasks without requiring paired images. Bao, et al. [30] proposed an innovative fusion architecture for handling visible and thermal imaging in few-shot learning scenarios.

 doi:10.20944/preprints202408.2054.v1

At the same time, many methods have been proposed to improve the effectiveness of multimodal image fusion. For example, TIRNet [31] focuses on addressing the redundancy problem associated with fused features, while RISNet [32] designed a new mutual information minimization module to reduce redundant information. Furthermore, PearlGAN [33] introduced a top-down guided attention module and structured gradient alignment loss to achieve hierarchical attention distribution, thereby reducing local semantic ambiguity and improving image encoding by leveraging contextual information.These research advancements demonstrate the diversity and cutting-edge nature of the multimodal image fusion detection field, offering new directions for further improving fusion effectiveness and detection performance.

2.3. Attention Mechanism

In deep learning, attention mechanisms have gained increasing importance as they enable networks to focus on key features by performing a series of transformations. This mechanism emphasizes the allocation of more information to the most significant feature representations while suppressing less useful ones. Self-attention methods achieve this by computing the weighted sum of all positions in an image to capture contextual information for a given position. SE [34] utilizes two fully connected layers to model the relationships between channels. NL (Spatial Attention) equips the model with spatial invariance, allowing the neural network to perform spatial transformations automatically while keeping feature maps invariant. ECA-Net [35] uses one-dimensional convolutional filters to generate channel weights, significantly reducing the complexity of the SE model. The non-local (NL) module proposed by Wang et al. [36] computes a correlation matrix between each spatial point in the feature map, producing a significant attention map. CBAM [37], GCNet [38], and SGE [39] sequentially combine spatial and channel attention, while Shuffle Attention (SA) and DANet [40] adaptively integrate local features with global dependencies by summing attention modules from different branches.

3. Methods

3.1. Overall Network Architecture

We designed the dual-branch object detection network IV-YOLO based on the YOLOv8 network. The overall structure of IV-YOLO is shown in Figure 1. In the backbone of the IV-YOLO object detection network, we employ dual branches to process visible light images and infrared images, respectively. Each branch's backbone consists of layers P1 to P5. The P1 layer uses a single-layer convolutional (Conv) structure, as shown in Equation (1). Here,$\mathbf{F}_{C_i} \in \mathbb{R}^{C_{in} \times H_{in} \times W_{in}}$,$\mathbf{F}_{C_i}$represents the input feature map, conv denotes a convolution operation with a kernel size of 3×3 and a stride of 2, BatchNorm2d is used to accelerate the network's convergence and enhance training stability, while the SiLU activation function ensures the non-linearity of the convolution. Layers P2 to P5 utilize Conv layers and the C2f structure from YOLOv8, with C2f illustrated in Figure 2.

$$\mathrm{Conv}\left(\mathbf{F}_{C_i}\right) = \mathrm{SiLU}(\mathrm{BatchNorm}\,2d(\mathrm{conv}_{3\times3,s=2}\ \left(\mathbf{F}_{C_i}\right))) \tag{1}$$

In the feature fusion section, we designed a novel bidirectional pyramid fusion (Bi-Fusion) module, the structure and characteristics of which are detailed in Section 3.2.1. In well-lit conditions, visible light images can capture rich details of target objects, whereas thermal infrared images exhibit significant contrast characteristics in low-light environments and can penetrate certain materials due to their thermal sensitivity. Therefore, we fused the features extracted by the two branches at different scales (P2, P3, P4, and P5) through the Bi-Fusion module, performing bidirectional pyramid feature fusion. The P2 layer, being the shallowest with the smallest receptive field, captures fusion features of small targets. The convolutional layers from P3 to P5 increase in depth, and their receptive fields progressively enlarge, thus obtaining fusion features for small, medium, and large targets, respectively. We found that the fused feature structure can effectively supplement the detail features of the infrared

 doi:10.20944/preprints202408.2054.v1

branch. Consequently, we fed the feature mapping vectors obtained from the fusion of P2 to P4 into the P3 to P5 layers of the infrared branch for further feature extraction.

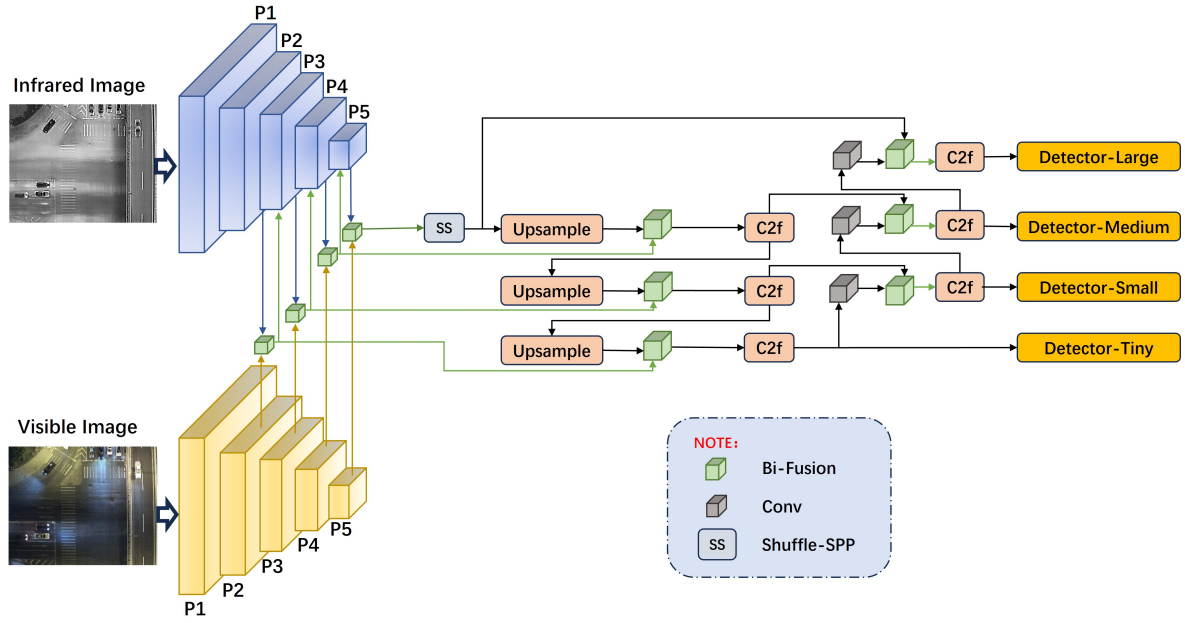

Figure 1. The overall architecture of the dual-branch IV-YOLO. This network is primarily used for detecting objects in complex environments. The backbone network simultaneously extracts features from both visible light and infrared images, and fuses the features extracted from P2-P5. The fused features are then passed to the infrared branch for further extraction. In the neck, we utilize the Shuffle-SPP structure, and the resulting features undergo a three-layer upsampling process, which integrates the features previously extracted by the backbone network, enabling the detection of objects at different scales.

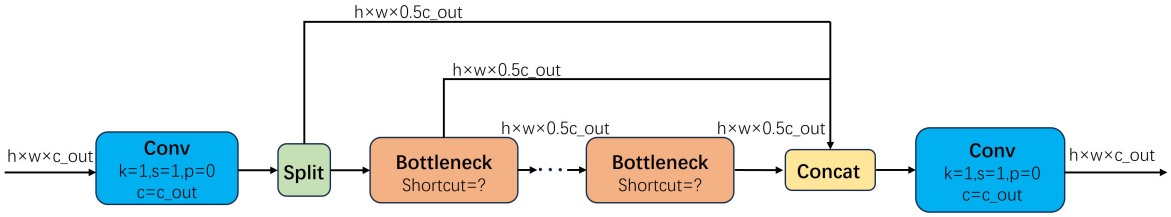

Figure 2. Diagram of the C2F Module.

In the design of the neck section, we introduced the Shuffle-SPP (Shuffle Spatial Pyramid Pooling) structure, which enhances feature utilization efficiency through multi-scale pooling and multi-layer convolution, combined with channel attention and spatial attention mechanisms, the structure and characteristics are detailed in Section 3.2.2. Additionally, drawing inspiration from the YOLOv8 architecture, we incorporated an upsampling layer and a convolution concatenation operation to preserve multi-scale features of the target. This design aims to mitigate the gradual loss of features caused by convolution operations, thereby enhancing the target detection capability.

Through the overall network design, this architecture can extract texture features and rich details from visible light images, as well as thermal radiation and material features from infrared images. Additionally, by combining multi-level feature fusion with the shuffle attention mechanism, it ensures the full utilization of both modalities' features, enhancing the network's robustness in complex environments.

 doi:10.20944/preprints202408.2054.v1

3.2. Feature Fusion Structure

Our feature fusion structure consists of two main components: the Bi-Directional Pyramid Feature Fusion Structure and the Shuffle Attention Spatial Pyramid Pooling Structure. The Bi-Directional Pyramid Feature Fusion Structure primarily integrates visible light image features with infrared image features, while the Shuffle Attention Spatial Pyramid Pooling Structure operates within the deeper layers of the network to fuse the extracted features, enhancing the connections between different channels. This allows the model to better focus on the channels or locations that contain target-specific features.

3.2.1. Bidirectional Pyramid Feature Fusion Structure

In the feature fusion module, we designed a bidirectional feature pyramid fusion structure called Bi-Fusion to effectively integrate the features of infrared and visible light images. The Bi-Fusion structure aims to enhance feature compensation and information interaction between the infrared and visible light channels, as illustrated in Figure 3. In the structure shown in Figure 3, we feed the features extracted from two branches into the main body of the fusion module. By establishing bidirectional connections between the two modalities, we ensure more efficient information flow and fusion of features at different scales. Additionally, we added extra edges between the input and output nodes at the same level and treated each bidirectional path as a feature network layer to optimize cross-scale connections. Moreover, we introduced learnable weights to ensure that Bi-Fusion effectively learns the features of different inputs, thereby enhancing the effectiveness of feature fusion.

The green module in Figure 4 illustrates the weight allocation structure, adding an extra weight to each input to enable the network to learn the importance of each input feature. We utilized Fast Normalized Fusion to assign weights to each feature, as shown in Equation (2).

$$O = \sum_i \frac{w_i}{\epsilon + \sum_j w_j} \cdot I_i \tag{2}$$

The condition $w_i \geqslant 0$ is ensured by applying a ReLU activation function after each w_i. The term $\epsilon = 0.0001$ is added to avoid division by zero and stabilize numerical computation. Equation (3-5) describes the weight parameter fusion of Bi-Fusion, as illustrated in Figure 3.

$$P^{td} = \text{Conv}\left(\frac{w_0 \cdot P_0^{\text{in}} + w_1 \cdot P_1^{\text{in}}}{w_0 + w_1 + \epsilon}\right) \tag{3}$$

$$P_0^{\text{out}} = \text{Conv}\left(\frac{(w_0 \cdot P_0^{\text{in}} + w_1 \cdot P^{td}) \cdot w_0}{(w_0 + w_1) \cdot w_0 + \epsilon}\right) \tag{4}$$

$$P_1^{\text{out}} = \text{Conv}\left(\frac{(w_1 \cdot P_1^{\text{in}} + w_0 \cdot P^{td}) \cdot w_1}{(w_1 + w_0) \cdot w_1 + \epsilon}\right) \tag{5}$$

Here, P_0 represents the features extracted from the infrared image, P_1 represents the features extracted from the visible light image, w_0 is the weight parameter for the infrared image features, and w_1 is the weight parameter for the visible light image features. P^{td} denotes the normalized weight parameter, improving the balance between accuracy and efficiency.

The Bi-directional Pyramid Feature Fusion Structure (Bi-Fusion) employs efficient bi-directional cross-scale connections and repeated module designs, allowing feature information to flow bi-directionally across different scales. This bi-directional information flow facilitates effective information exchange between different scales. The purpose of this design is to enhance the efficiency and effectiveness of feature fusion by strengthening bi-directional feature flow, thereby improving target detection performance.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 August 2024 doi:10.20944/preprints202408.2054.v1

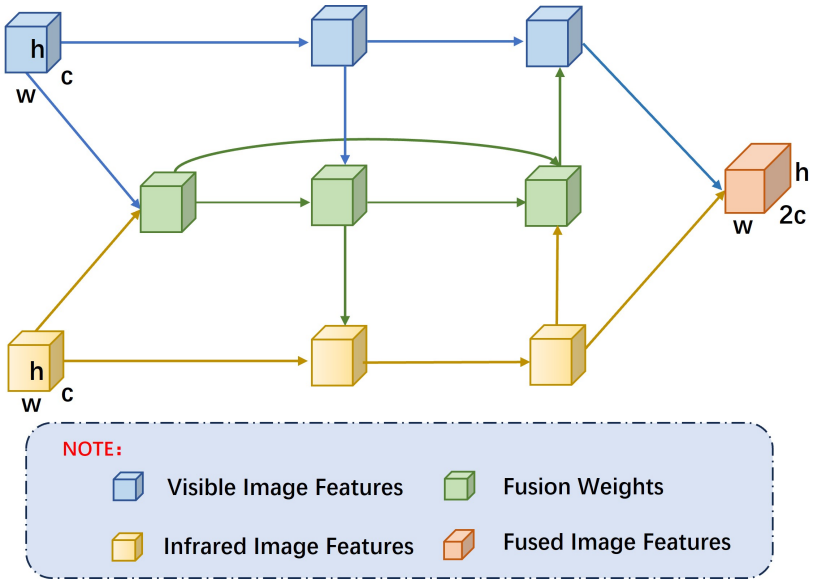

Figure 3. Bidirectional Pyramid Feature Fusion Structure, Bi-Fusion Structure.

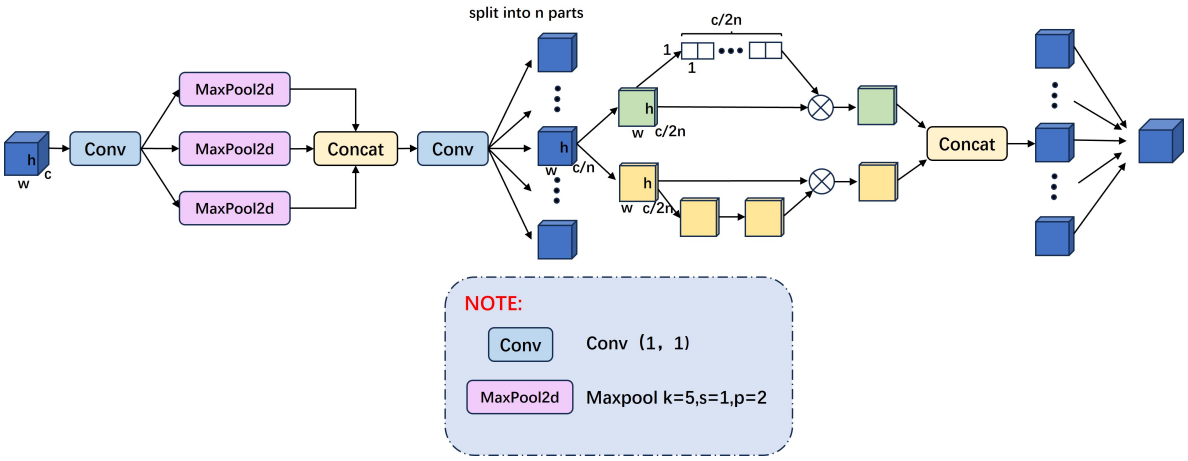

Figure 4. Shuffle Attention Spatial Pyramid Pooling Structure.

3.2.2. Shuffle Attention Spatial Pyramid Pooling Structure

To enable deep networks to learn more mixed features from infrared and visible light images, thereby making better use of the characteristics of both modes, we adapted and improved the shuffle module design from SA-NET [41]. After feature extraction and fusion through layers P1 to P5, we input the deepest fused features into the Shuffle-SPP module. This module facilitates information exchange between the two modal images through multi-scale pooling, convolution concatenation, and the shuffle attention spatial pyramid pooling structure. By combining efficient layer aggregation, it enhances the model's expressive capability, enabling it to extract richer features. The overall structure of the Shuffle-SPP module is shown in Figure 4.

In Figure 4, we first adjust the number of channels using a Conv layer, followed by three MaxPool2d layers for multi-scale feature extraction. Next, we aggregate the extracted spatial information and restore the features through a Conv layer to maintain the original image dimensions. To better integrate the different characteristics of infrared and visible light modes, we divide the channels into n sub-features X_k and iterate through each one. Each sub-feature X_k is further divided into two parts: X_{k1} and X_{k2}. The X_{k1} part fully utilizes different channels and applies the channel attention mechanism to focus on key features, while the X_{k2} part leverages the spatial relationships of the features, using the spatial attention mechanism to highlight important features.

 doi:10.20944/preprints202408.2054.v1

The design of the X_{k1} part references the SE module and employs a channel attention mechanism. The module uses a single-layer transformation of GAP+Scale+Sigmoid to convert the feature X_{k1} from dimensions $\frac{w \cdot h \cdot c}{2n}$ to 11c through global average pooling, as shown in equation (6) and (7). Then, the obtained channel weights are normalized and multiplied with X_{k1} to obtain the feature module X'_{k1} with channel attention. As training progresses, the module gradually captures specific semantic information.

$$s = \mathcal{F}_{gp}(X_{k1}) = \frac{1}{h \cdot w} \sum_{i=1}^{h} \sum_{j=1}^{w} X_{k1}(i, j) \tag{6}$$

$$X'_{k1} = \sigma(s) \cdot X_{k1} \tag{7}$$

The design of the X_{k2} part employs a spatial attention mechanism, which, unlike channel attention, focuses on "where" the information is, complementing channel attention. First, we use Group Norm (GN) [42] on X_{k2} to obtain spatial statistics and apply a function to enhance the representation of X'_{k2}. Subsequently, the obtained spatial weights are multiplied with X_{k2} to produce the feature module X'_{k2} with spatial attention ,as shown in equation (8). As training progresses, the module gradually captures specific spatial information.

$$X'_{k2} = \sigma(W_2 \cdot GN(X_{k2}) + b_2) \cdot X_{k2} \tag{8}$$

Then, all sub-features are aggregated to obtain a feature module with comprehensive information interaction.

The Shuffle-SPP module achieves global context information acquisition by rearranging and pooling features across channel and spatial dimensions. It effectively combines relevant information from different channels while capturing broader spatial context. This process allows the network to focus more precisely on the semantic and positional information of the target when processing deep features, thereby enhancing the model's expressiveness and the accuracy of target detection. In this way, the Shuffle-SPP module not only enriches feature representation but also strengthens the network's ability to recognize targets in complex scenarios.

3.3. Loss Function

By analyzing the dataset, we categorize the detection targets into four types: tiny, small, medium, and large. However, there is an imbalance in the sample quantities among these categories, with more medium and large targets and fewer tiny and small targets. Medium and large targets generally benefit from sufficient feature extraction, making them easier to locate and detect, whereas tiny and small targets are harder to accurately locate. Therefore, we focus particularly on the performance of these difficult-to-detect samples in bounding box regression and improve our loss function by referencing PIoU v2.

To better handle the regression samples of different categories and focus on their respective detection tasks, we adopt a linear interval mapping method to reconstruct the IoU loss. This method enhances the regression accuracy for edge targets. The formula is as follows:

$$IoU^{\text{focaler}} = \begin{cases} 0, & IoU < d \\ \frac{IoU-d}{u-d}, & d \leqslant IoU \leqslant u \\ 1, & IoU > u \end{cases} \tag{9}$$

Here, IoU^{focaler} represents the reconstructed Focaler-IoU, and IoU is the original value, $[d, u] \in [0, 1]$. By adjusting the values of d and u, we can make IoU^{focaler} focus on different regression samples. The loss is defined as follows:

 doi:10.20944/preprints202408.2054.v1

$$L_{\text{Focaler-IoU}} = 1 - IoU^{\text{focaler}} \tag{10}$$

By optimizing $L_{\text{Focaler-IoU}}$, the parameters of the feature extraction network can be improved, redundant image features can be eliminated, and the network's generalization ability can be enhanced, thereby speeding up convergence. For coordinate location errors, we choose to use Complete IoU (CIoU) Loss as the loss function to ensure more stable bounding box regression, as shown in equation (11).

$$L_{\text{Focaler-CIoU}} = L_{CIoU} + IoU - IoU^{\text{Focaler}} \tag{11}$$

PIoU v2 requires only one hyperparameter, simplifying the tuning process. It combines a non-monotonic attention function controlled by a single hyperparameter with PIoU. The PIoU v2 loss is derived by adding an attention layer to the PIoU loss, enhancing the focus on medium to high-quality anchor boxes and improving the performance of the object detector. By emphasizing the optimization of anchor boxes with moderate quality during the regression process, PIoU v2 ensures better detection results.

4. Results

In this chapter, we provide a detailed description of the implementation of the infrared and visible light dual-branch image detection network, along with the hardware and software configuration details. To verify the effectiveness of the proposed method, we conducted extensive experiments on several public datasets, including the Drone Vehicle dataset, the KAIST dataset, and the FLIR pedestrian dataset. The experimental results demonstrate that our method performs excellently across these datasets.

4.1. Dataset Introduction

4.1.1. Drone Vehicle Dataset

The Drone Vehicle dataset [14] is a public dataset specifically designed for UAV image target detection and classification tasks, widely used in traffic monitoring and intelligent transportation systems research. Captured by drones, the dataset provides high-resolution aerial images covering various traffic scenes such as urban roads, highways, and parking lots. It includes detailed annotations of the locations, bounding boxes, and class labels of various vehicles, including cars, trucks, and buses.

The images span different environmental conditions from day to night, including both infrared and visible light images. The entire dataset comprises 15,532 pairs of images (a total of 31,064 images) and 441,642 annotated instances. Additionally, the dataset includes real-world occlusions and scale variations.

4.1.2. FLIR Dataset

The FLIR dataset [15], released by FLIR Systems, is a public dataset widely used for infrared image object detection and pedestrian detection research. This dataset primarily includes infrared images and provides corresponding visible light images, enabling researchers to explore multimodal fusion methods. The FLIR dataset covers various scenes and environmental conditions, including day, night, urban streets, and rural roads, helping to evaluate the robustness of detection algorithms under different lighting and complex backgrounds.

The dataset contains over 10,000 pairs of 8-bit infrared images and 24-bit visible light images, featuring objects such as people, vehicles, and bicycles in both daytime and nighttime scenes. The resolution of the infrared images is 640 × 512, while the resolution of the visible light images ranges from 720 × 480 to 2048 × 1536.

 doi:10.20944/preprints202408.2054.v1

4.1.3. KAIST Dataset

The KAIST dataset [13], released by the Korea Advanced Institute of Science and Technology (KAIST), is a public dataset widely used for multimodal object detection and tracking research. It includes both visible light images and infrared images, making it particularly suitable for studying the fusion of information from different modalities to enhance detection performance. The KAIST dataset covers various scenes and environmental conditions, including day, night, clear, and rainy weather, to help evaluate the robustness of algorithms under different lighting and weather conditions. This dataset specifically focuses on pedestrian detection tasks and provides detailed annotation information, such as the position, size, and category labels of pedestrians.

The dataset is divided into 12 subsets, with sets 00 to 05 serving as training data (sets 00 to 02 for daytime scenes, sets 03 to 05 for nighttime scenes) and sets 06 to 11 as testing data (sets 06 to 08 for daytime scenes, sets 09 to 11 for nighttime scenes). The image resolution is 640 × 512, and the dataset includes a total of 95,328 images, with each image containing both visible light and infrared versions. The KAIST dataset covers various common traffic scenes during both day and night, including campuses, streets, and rural areas, and provides 103,108 densely annotated instances.

In our experiments, we resized each visible image to 640 × 640. Table 1 summarizes the dataset information used for training and testing.

Table 1. This is a table caption. Tables should be placed in the main text near to the first time they are cited.

Hyper Parameter	Drone Vehicle Dataset	FLIR Dataset	KAIST Dataset
Scenario	drone	adas	pedestrian
Modality	Infrared + Visible	Infrared + Visible	Infrared + Visible
Images	56878	14000	95328
Categories	5	4	3
Labels	190.6K	14.5K	103.1K
Resolution	840 × 712	640 × 512	640 × 512

4.2. Implementation Details

We chose the YOLOv8 network as the primary framework. During data augmentation, each image had a 50% chance of being randomly flipped horizontally, and mosaic operation was employed to stitch multiple images into one, increasing the complexity and diversity of training samples. The entire network was optimized using the AdamW optimizer with weight decay, training for 300 epochs. The learning rate was set to 0.001, and the batch size was 32. Weight decay was set to 0.0005, and warmup momentum was set to 0.8. The AdamW optimizer combines the advantages of the Adam optimizer while effectively reducing the risk of overfitting through weight decay, leading to more stable and efficient training in most cases.

We implemented our code using the PyTorch framework and conducted experiments on a workstation equipped with an NVIDIA GTX4090 GPU. We summarize the experimental environment and parameter settings in Table 2. The hyperparameters for the datasets used in this paper are shown in Table 3. The numbers of infrared and visible light images were equal. When using these datasets for network training and testing, we performed data cleaning operations.

 doi:10.20944/preprints202408.2054.v1

Table 2. Experimental Setup and Parameter Settings.

Category	Parameter
CPU Intel	i7-12700H
GPU	RTX 4090
System	Windows11
Python	3.8.19
PyTorch	1.12.1
Training Epochs	300
Learning Rate	0.01
Weight Decay	0.0005
Momentum	0.937

Table 3. The hyperparameters of the datasets used in this paper. 'test-val' indicates that the same dataset is used for both testing and validation in this study.

Hyper Parameter	DroneVehicle Dataset	FLIR Dataset	KAIST Dataset
Visible Image Size	640 × 640	640 × 640	640 × 640
Infrared Image Size	640 × 640	640 × 640	640 × 640
Visible Image	28439	8437	8995
Infrared Image	28439	8437	8995
Training set	17990	7381	7595
Validation set	1469	1056	1400
Testing set	8980	1056 (test-val)	1400 (test-val)

4.3. Evaluation Metrics

Experiment on the DroneVehicle Dataset

We use Precision, Recall, and mean Average Precision (mAP) to evaluate the detection performance of different methods. In the experiments presented in this paper, we primarily use the values of precision and recall to measure the performance of the network, as calculated in Equations (12) and (13).

$$precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{12}$$

$$recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{13}$$

TP (True Positive) represents the number of samples correctly predicted as positive by the model. FP (False Positive) represents the number of samples incorrectly predicted as positive by the model. FN (False Negative) represents the number of positive samples incorrectly predicted as negative by the model.

AP (Average Precision) refers to the area under the P-R curve enclosed by the coordinates. The closer the AP value is to 1, the better the detection performance of the algorithm. The calculation process of AP can be summarized as follows:

$$\mathrm{AP} = \int_0^1 p(r)dr \tag{14}$$

mAP represents the average of the AP for each class. By considering both precision and recall, it provides a balanced measure of model performance. It is especially important in multi-class detection tasks as it ensures that the model performs well across all categories. Additionally, mAP is robust to class instance imbalances, providing a fair comparison of models. Therefore, mAP is widely used to evaluate the performance of multi-class object detection tasks and is utilized in our experiments to assess detection accuracy.The formula for calculating mAP is as follows:

 doi:10.20944/preprints202408.2054.v1

$$\text{mAP} = \frac{1}{C}\sum_{c=1}^{C}\text{AP} \tag{15}$$

4.4. Analysis of Results

In this section, we test the IV-YOLO network on three test datasets and compare the detection results with those of advanced methods.

4.4.1. Experiment on the DroneVehicle Dataset

To validate the effectiveness of our proposed dual-branch object detection method for detecting small targets under complex conditions, we conducted a series of experiments on the Drone Vehicle dataset. In these experiments, we preprocessed the dataset by cropping 100 pixels of white border from each side of the images, resulting in images of size 640×512 pixels, and modified the detection head to use a Rotated Object Detection Head (OBB). Furthermore, in the Drone Vehicle dataset, the shapes of Freight Cars and Vans are very similar, and many existing detection methods typically ignore these two classes to avoid errors due to fine-grained classification. In contrast, we used the complete Drone Vehicle dataset in our experiments to assess our network's ability to extract and fuse fine-grained features from dual modalities. The performance comparison between our method and other networks is shown in Table 4.

Since many network models are designed for single-modal image detection, we evaluated these models on both visible light and infrared images. As shown in Table 4, the YOLOv8 model performs exceptionally well in terms of detection accuracy and also demonstrates high detection speed in single-modal scenarios. This is why we chose YOLOv8 as the core framework for the IV-YOLO algorithm. Through innovative improvements, our proposed IV-YOLO algorithm achieved an accuracy of 74.6% on the Drone Vehicle dataset, the highest among all networks tested. Specifically, for Freight Cars and Vans, which are targets that are difficult for other networks to distinguish, our network achieved the highest detection accuracies of 63.1% and 53% respectively by extracting and fusing fine-grained features. The results in Table 4 indicate that our network not only effectively integrates features from both modalities for robust detection in complex environments but also enhances detection performance for visually similar targets through dual-branch feature fusion.

Table 4. Evaluation results based on the Drone Vehicle dataset. All values are expressed as percentages. The top-ranked results are highlighted in green.

Method	Modality	Car	Freight Car	Truck	Bus	Van	mAP
RetinaNet(OBB)[43]	Visible	67.5	13.7	28.2	62.1	19.3	38.2
Faster R-CNN(OBB)[17]	Visible	67.9	26.3	38.6	67.0	23.2	44.6
Faster R-CNN(Dpool)[44]	Visible	68.2	26.4	38.7	69.1	26.4	45.8
Mask R-CNN[45]	Visible	68.5	26.8	39.8	66.8	25.4	45.5
Cascade Mask R-CNN[46]	Visible	68.0	27.3	44.7	69.3	29.8	47.8
RoITransformer[2]	Visible	68.1	29.1	44.2	70.6	27.6	47.9
YOLOv8(OBB)[12]	Visible	97.5	42.8	62.0	94.3	46.6	68.6
RetinaNet(OBB)[43]	Infrared	79.9	28.1	32.8	67.3	16.4	44.9
Faster R-CNN(OBB)[17]	Infrared	88.6	35.2	42.5	77.9	28.5	54.6
Faster R-CNN(Dpool)[44]	Infrared	88.9	36.8	47.9	78.3	32.8	56.9
Mask R-CNN[45]	Infrared	88.8	36.6	48.9	78.4	32.2	57.0
Cascade Mask R-CNN[46]	Infrared	81.0	39.0	47.2	79.3	33.0	55.9
RoITransformer[2]	Infrared	88.9	41.5	51.5	79.5	34.4	59.2
YOLOv8(OBB)[12]	Infrared	97.1	38.5	65.2	94.5	45.2	68.1
UA-CMDet[14]	Visible+Infrared	87.5	46.8	60.7	87.1	38.0	64.0
Dual-YOLO[7]	Visible+Infrared	98.1	52.9	65.7	95.8	46.6	71.8
IR-YOLO(Ours)	Visible+Infrared	97.2	63.1	65.4	94.3	53.0	74.6

 doi:10.20944/preprints202408.2054.v1

Figure 5 presents the visualization results of target detection on the Drone Vehicle dataset. The visualized images are organized into six groups, each consisting of three rows. In each group, the left side shows the detection results for visible light images, while the right side displays the results for infrared images. The first row demonstrates that our network can achieve robust target detection even under low-light conditions or with soft occlusions. The second row indicates that the network effectively extracts fine-grained features, allowing it to distinguish between visually similar targets. The third row shows that the network maintains strong robustness when dealing with densely populated target scenes.

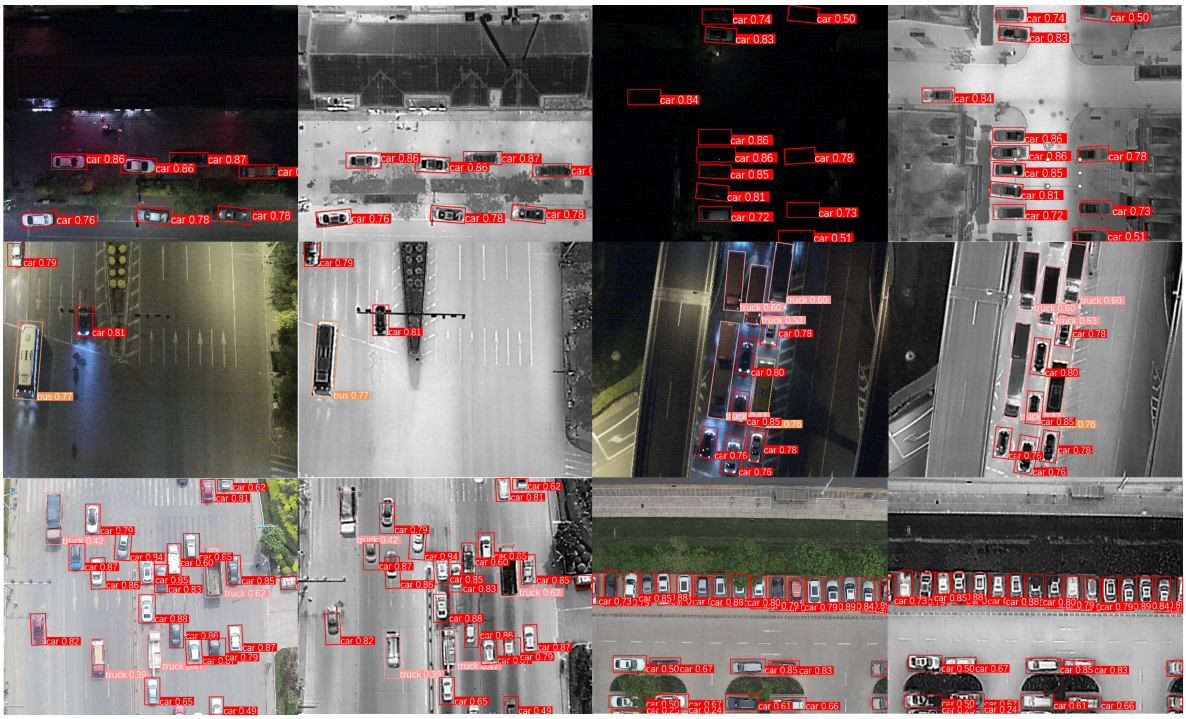

Figure 5. Visualization of IV-YOLO Detection Results Based on the Drone Vehicle Dataset.

4.4.2. Experiments on the FLIR Dataset

We conducted a series of experiments on the FLIR dataset to validate the effectiveness of the proposed method. Additionally, we compared the performance of common methods such as SSD, RetinaNet, YOLOv5s, and YOLOF, along with the Dual-YOLO network, which has shown high performance on the FLIR dataset. The final experimental results are presented in Table 5.

The experimental results on the FLIR dataset are presented in Table 6. As shown in the table, our network achieved the highest mAP value on this dataset compared to other methods. By utilizing multi-scale feature fusion and three upsampling operations in the neck, our network enhanced its ability to extract features from small objects. Table 6 indicates a significant improvement in the detection accuracy of small objects like bicycles, which are typically challenging to identify. Although there was a slight decline in performance when detecting large objects such as cars and pedestrians, likely due to a reduced capability in capturing global features, the overall mAP results demonstrate that our network effectively extracts multi-scale features, particularly in layers with smaller receptive fields, leading to improved detection of small objects. Furthermore, our network's fusion module successfully extracts and integrates both shared and unique features from the two modalities. Through the fusion of weighted parameters, the two modalities mutually enhance and complement each other, resulting in superior detection performance.

 doi:10.20944/preprints202408.2054.v1

Table 5. The object detection results on the FLIR dataset, calculated at a single IoU threshold of 0.5. All values are expressed as percentages, with the top-ranked results highlighted in green.

Method	Person	Bicycle	Car	mAP
Faster R-CNN[17]	39.6	54.7	67.6	53.9
SSD[16]	40.9	43.6	61.6	48.7
RetinaNet[43]	52.3	61.3	71.5	61.7
FCOS[47]	69.7	67.4	79.7	72.3
MMTOD-UNIT[14]	49.4	64.4	70.7	61.5
MMTOD-CG[14]	50.3	63.3	70.6	61.4
RefineDet[48]	77.2	57.2	84.5	72.9
TermalDet[14]	78.2	60.0	85.5	74.6
YOLO-FIR[49]	85.2	70.7	84.3	80.1
YOLOv3-tiny[20]	67.1	50.3	81.2	66.2
IARet[50]	77.2	48.7	85.8	70.7
CMPD[51]	69.6	59.8	78.1	69.3
PearlGAN[33]	54.0	23.0	75.5	50.8
Cascade R-CNN[46]	77.3	84.3	79.8	80.5
YOLOv5s[22]	68.3	67.1	80.0	71.8
YOLOF[52]	67.8	68.1	79.4	71.8
Dual-YOLO[7]	88.6	66.7	93	84.5
IV-YOLO(Ours)	86.6	77.8	92.4	85.6

Figure 6 illustrates the visualized results of target detection on the FLIR dataset. The images in the first row demonstrate that, regardless of whether the background lighting is overly bright or too dim, our network is able to accurately locate and detect targets by fusing features from two modalities. The second row of images indicates that the network effectively extracts fine-grained features of objects at different scales through its dual-branch structure, and, combined with a specially designed loss function, accurately detects each target even in the presence of occlusions and densely packed objects. The third row showcases the network's strength in multi-scale feature extraction, particularly in capturing more subtle details at smaller receptive fields, thereby significantly improving the detection of small targets.

Figure 6. Visualization of IV-YOLO Detection Results on the FLIR Dataset.

 doi:10.20944/preprints202408.2054.v1

Table 6. The object detection results on the FLIR dataset, calculated at a single IoU threshold of 0.5. All values are expressed as percentages, with the top-ranked results highlighted in green.

Method	Precision	Recall	mAP
ForkGAN[53]	33.9	4.6	4.9
ToDayGAN[54]	11.4	14.9	5.0
UNIT[55]	40.9	43.6	11.0
PearlGAN[33]	21.0	39.8	25.8
Dual-YOLO[7]	75.1	66.7	73.2
IV-YOLO(OURS)	77.2	84.5	75.4

4.4.3. Experiments Based on the KAIST Dataset

To further validate the effectiveness and robustness of our proposed IV-YOLO algorithm, we conducted experiments on the challenging KAIST dataset. Given the presence of numerous blank and misaligned images in the KAIST dataset, we first performed data cleaning and then carried out the object detection task on the processed dataset. After comparing our results with several popular methods, our experimental outcomes are presented in Table 6.

From Table 6, it is evident that compared to the network we proposed for the KAIST dataset, the accuracy of the first three methods is significantly lower. This discrepancy is mainly due to the lack of data preprocessing during the training on the KAIST dataset, which led to issues such as blank images and misaligned labels, thereby affecting the accuracy of the results. After preprocessing the dataset, we compared our network with the state-of-the-art Dual-YOLO and PearlGAN. The results are shown in the last three rows of Table 6, which demonstrate that our method achieved the best performance in terms of Precision, Recall, and mean Average Precision (mAP).

Figure 7 presents the visualization of our test results along with some data from the KAIST dataset. The figure includes 6 sets, with a total of 12 images. From the scenes in the first row, it is evident that our network demonstrates strong robustness to scale variations and changes in image brightness. In the second row, our IV-YOLO is able to accurately detect pedestrians on the street despite overlapping and occlusions. The scenes in the third row indicate that through feature fusion, our network can accurately identify targets even in complex backgrounds.

Figure 7. Visualization of IV-YOLO Detection Results on the KAIST Pedestrian Dataset.

 doi:10.20944/preprints202408.2054.v1

4.5. Parameter Analysis

In this study, we evaluated the parameter efficiency and inference speed of the proposed model on the Drone Vehicle and FLIR datasets. To ensure the accuracy of the experimental results, all tests for parameter efficiency and inference speed were conducted on an NVIDIA GeForce RTX 3090 GPU, with the parameter size represented in MB using 16-bit precision.

As shown in Table 7, our IV-YOLO network maintains high detection accuracy while keeping the parameter size low. This model can be flexibly deployed on devices with limited resources while still delivering exceptional performance. When running on the NVIDIA GeForce RTX 3090 GPU, IV-YOLO achieves a real-time processing capability of up to 198 frames per second (FPS). This high FPS demonstrates the model's ability to perform fast and accurate object detection with high-resolution input, making it particularly suitable for applications where speed and quality are critical.

Table 7. Model complexity and runtime comparison of IV-YOLO and the plain counterparts.

Method	Dataset	Params	Runtime(fps)
Faster R-CNN(OBB)[17]	Drone Vehicle	58.3M	5.3
Faster R-CNN(Dpool)[44]	Drone Vehicle	59.9M	4.3
Mask R-CNN[45]	Drone Vehicle	242.0M	13.5
RetinaNet[44]	Drone Vehicle	145.0M	15.0
Cascade Mask R-CNN[46]	Drone Vehicle	368.0M	9.8
RolTransformer[2]	Drone Vehicle	273.0M	7.1
YOLOv7[24]	Drone Vehicle	72.1M	161.0
YOLOv8n[12]	Drone Vehicle	5.92M	188.6
SSD[16]	FLIR	131.0M	43.7
FCOS[47]	FLIR	123.0M	22.9
RefineDet[48]	FLIR	128.0M	24.1
YOLO-FIR[49]	FLIR	7.1M	83.3
YOLOv3-tiny[20]	FLIR	17.0M	66.2
Cascade R-CNN[46]	FLIR	165.0M	16.1
YOLOv5s[22]	FLIR	14.0M	41.0
YOLOF[52]	FLIR	44.0M	32.0
Dual-YOLO[7]	FLIR	175.1M	62.0
IV-YOLO(Ours)	Drone Vehicle/FLIR	4.31M	208.2

4.6. Ablation Study

To gain a deeper understanding of the practical roles of various modules and components in the proposed model, we designed and conducted a series of ablation experiments. These experiments aim to remove key components of the network and observe their impact on overall performance to clarify the contribution of each module to the model. To eliminate the impact of different IoU settings on the experiment, we used both mAP@0.5 and mAP@0.5:0.95 as evaluation metrics to assess accuracy in this experiment.

First, we performed an ablation experiment on the feature fusion module by removing the Shuffle-SPP from the overall model. The Shuffle-SPP was initially designed to enhance the internal correlation between deep features, improving the deep network's ability to recognize and locate targets. After removing this component and re-running the experiments on the standard dataset, we found that the model's mAP dropped by 1.5%. This decline clearly indicates that incorporating the Shuffle-SPP structure improves the model's performance in deep feature extraction.

Next, we conducted a similar ablation experiment on the Bi-Fusion module. Compared to the removal of Shuffle-SPP, removing Bi-Fusion had a relatively larger impact on overall model performance, resulting in the mAP dropping to 82.9%. This decrease demonstrates that the Bi-Fusion structure effectively merges the features of visible light images and infrared images. Given its broader range of influence, its impact on the results is also relatively significant. Proper use of this structure can effectively enhance the model's performance.

 doi:10.20944/preprints202408.2054.v1

Table 8. Ablation experiments of the feature fusion structure, all values are expressed as a percentage.

Shuffle-SPP	Bi-concat	Person	Bicycle	Car	mAP@0.5	mAP@0.5:0.95
×	✓	83.7	76.4	92.1	84.1	50.1
✓	×	83.2	73.8	91.6	82.9	49.1
✓	✓	86.6	77.8	92.4	85.6	51.2

We also removed an upsampling layer and a concatenation structure from the overall network, which were specifically optimized for small object detection. The small object detection network was designed to enhance the model's ability to detect small objects. After removing this component, the model's performance in small object detection exhibited significant fluctuations, and it also showed noticeable degradation when detecting objects with high similarity.

To more comprehensively assess the contribution of the additional structures, we removed the three components simultaneously and analyzed their combined impact on the model. The final ablation experiment results are shown in Figures 8 and 9. In these figures, 'A' represents our IV-YOLO network, 'N' denotes the Shuffle-SPP structure, 'F' stands for the Bi-Fusion structure, and 'E' indicates the modified network designed to enhance small object detection capabilities.

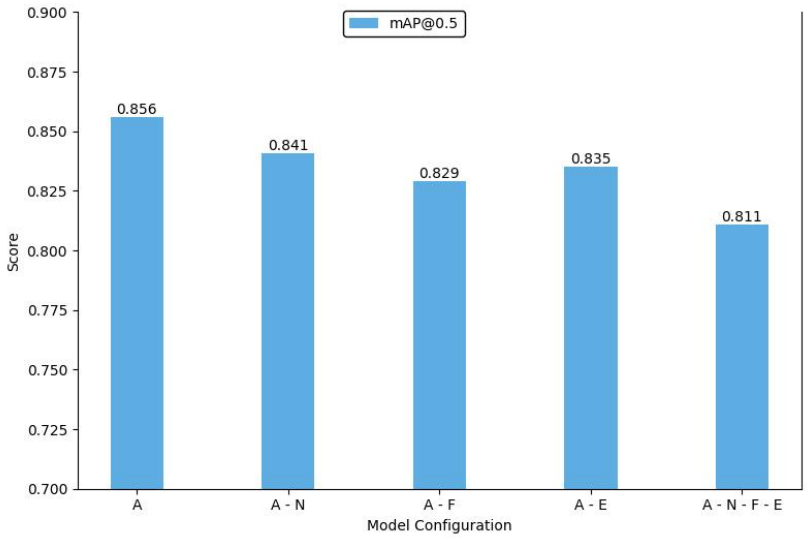

Figure 8. Bar chart of mAP@0.5 in the ablation experiments.

In conclusion, the ablation experiments clearly demonstrate the critical roles of the various modules and components in our proposed model. By analyzing these components individually, we validated the significant contributions of each module to specific tasks, thereby proving the rationality and effectiveness of the model's structural design. The experimental results indicate that these modules exhibit significant complementarity across different tasks and scenarios. Integrating these modules maximizes the model's performance, while the absence of any one module would significantly diminish the overall effectiveness of the model.

 doi:10.20944/preprints202408.2054.v1

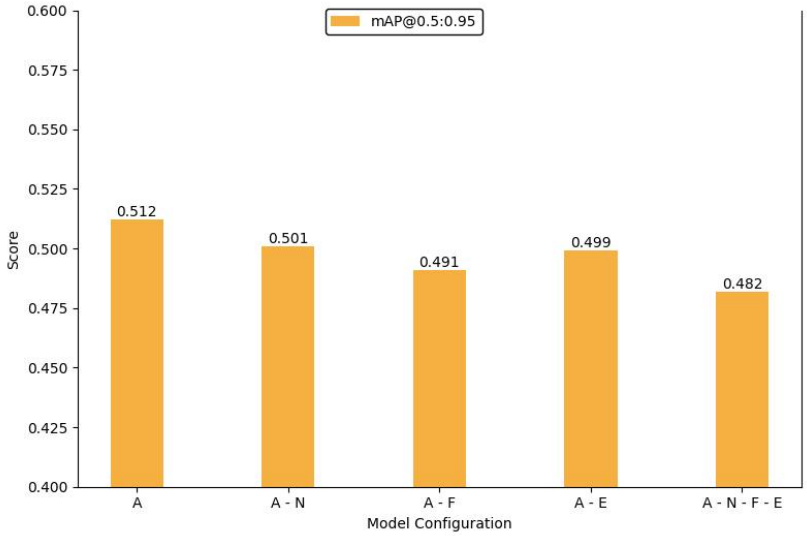

Figure 9. Bar chart of mAP@0.5:0.95 in the ablation experiments.

5. Conclusions

This paper presents a dual-branch target detection network based on YOLOv8 that fuses infrared and visible light images, named IV-YOLO. The network is designed to fully extract features from both modalities during the dual-branch feature extraction process while enabling target detection at multiple scales. Additionally, we have developed a bidirectional pyramid feature fusion structure (Bi-Fusion), which fuses features from different modalities across multiple scales using weighted parameters. This approach effectively reduces the number of parameters and avoids feature redundancy, thereby enhancing the network's fusion performance. To further improve the network's expressive power, we introduce the Shuffle-SPP module, which incorporates shuffling attention and spatial pyramid pooling to capture global context information across channels and spatial dimensions. This allows the network to focus more on the semantic and positional information of the target in deep features, significantly enhancing the model's expressiveness. Finally, by integrating the PIoU v2 loss function, we accelerate the convergence of the detection box for targets of different sizes, improving the fitting accuracy for small targets and speeding up the overall network convergence.

Experimental results show that our IV-YOLO network achieves a mAP of 74.6% on the Drone Vehicle dataset, 75.4% on the KAIST dataset, and 85.6% on the FLIR dataset, demonstrating excellent performance in feature extraction, fusion, and target detection for both infrared and visible light images. More importantly, the network has a parameter size of only 4.31M, significantly lower than other networks, showcasing its potential for deployment on mobile devices. However, despite the good performance of IV-YOLO in dual-modal target detection tasks, there are still some limitations:

(1)Further optimization of speed and memory utilization is required before hardware deployment.

(2)The availability of dual-modal datasets is currently limited, and there is a lack of diverse scenarios and diverse targets. Therefore, the effectiveness of our method requires further research and validation.

Although single-modal detection performs well in specific tasks, its limitations become apparent when dealing with complex and variable real-world scenarios. In contrast, dual-modal detection, by leveraging the advantages of multiple data sources, provides richer feature information and significantly improves the accuracy and robustness of target detection. Therefore, dual-modal detection technology is undoubtedly the main direction for future advancements in target detection, offering strong support for addressing more complex detection tasks.

 doi:10.20944/preprints202408.2054.v1

Author Contributions: Conceptualization, X.Y. ; methodology, X.Y., D.Z. and D.T.; software, X.Y.; validation, X.Y. and W.Z.; formal analysis, C.W.; investigation, D.T.; resources, X.Y. and D.Z; data curation, X.Y. and W.Z.; writing—original draft preparation, X.Y.; writing—review and editing, X.Y., D.Z. and D.T.; visualization, X.Y.; supervision, W.Z. All authors have read and agreed to the published version of the manuscript.

Funding: Not applicable.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The DroneVehicle remote sensing dataset is obtained from https://github.com/VisDrone/DroneVehicle, accessed on 29 December 2021. The KAIST pedestrian dataset is obtained from https://github.com/SoonminHwang/rgbt-ped-detection/tree/master/data, accessed on 12 November 2021. The FLIR dataset is obtained from https://www.flir.com/oem/adas/adasdataset-form/, accessed on 19 January 2022.

Acknowledgments: Not applicable.

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

References

1. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. Proceedings of the AAAI conference on artificial intelligence, 2017, Vol. 31.
2. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning RoI transformer for oriented object detection in aerial images. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 2849–2858.
3. Huang, X.; Liu, M.Y.; Belongie, S.; Kautz, J. Multimodal unsupervised image-to-image translation. Proceedings of the European conference on computer vision (ECCV), 2018, pp. 172–189.
4. Holst, G.C. Infrared Imaging Systems: Design, Analysis, Modeling, and Testing. *Infrared Imaging Systems: Design, Analysis, Modeling, and Testing* **1990**, *1309*.
5. Ma, K.; Yeganeh, H.; Zeng, K.; Wang, Z. High dynamic range image compression by optimizing tone mapped image quality index. *IEEE Transactions on Image Processing* **2015**, *24*, 3086–3097.
6. French, G.; Finlayson, G.; Mackiewicz, M. Multi-spectral pedestrian detection via image fusion and deep neural networks. *Journal of Imaging Science and Technology* **2018**, pp. 176–181.
7. Bao, C.; Cao, J.; Hao, Q.; Cheng, Y.; Ning, Y.; Zhao, T. Dual-YOLO architecture from infrared and visible images for object detection. *Sensors* **2023**, *23*, 2934.
8. Guan, D.; Cao, Y.; Yang, J.; Cao, Y.; Yang, M.Y. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion* **2019**, *50*, 148–157.
9. Kim, J.U.; Park, S.; Ro, Y.M. Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology* **2021**, *32*, 1510–1523.
10. Ma, J.; Yu, W.; Liang, P.; Li, C.; Jiang, J. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Information fusion* **2019**, *48*, 11–26.
11. Zhao, Z.; Xu, S.; Zhang, J.; Liang, C.; Zhang, C.; Liu, J. Efficient and model-based infrared and visible image fusion via algorithm unrolling. *IEEE Transactions on Circuits and Systems for Video Technology* **2021**, *32*, 1186–1196.
12. Ultralytics. YOLOv8. https://github.com/ultralytics/ultralytics, 2023.
13. Hwang, S.; Park, J.; Kim, N.; Choi, Y.; So Kweon, I. Multispectral pedestrian detection: Benchmark dataset and baseline. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1037–1045.
14. Sun, Y.; Cao, B.; Zhu, P.; Hu, Q. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology* **2022**, *32*, 6700–6713.
15. FLIR ADAS Dataset. https://www.flir.com/oem/adas/adas-dataset-form, 2022. Accessed: 2022-01-19.
16. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer, 2016, pp. 21–37.
17. Girshick, R. Fast r-cnn. Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

 doi:10.20944/preprints202408.2054.v1

18. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788. doi:10.1109/CVPR.2016.91.
19. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6517–6525. doi:10.1109/CVPR.2017.690.
20. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv preprint arXiv:1804.02767* **2018**.
21. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv preprint arXiv:2004.10934* **2020**.
22. Jocher, G. YOLOv5 by Ultralytics. https://github.com/ultralytics/yolov5, 2020.
23. Li, C.Y.; Wang, N.; Mu, Y.Q.; Wang, J.; Liao, H.Y.M. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. *arXiv preprint arXiv:2209.02976* **2022**.
24. 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. *arXiv preprint arXiv:2207.02696* **2022**.
25. Wang, C.Y.; Yeh, I.H.; Liao, H.Y.M. Yolov9: Learning what you want to learn using programmable gradient information. *arXiv preprint arXiv:2402.13616* **2024**.
26. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. Yolov10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458* **2024**.
27. Farahnakian, F.; Heikkonen, J. Deep learning based multi-modal fusion architectures for maritime vessel detection. *Remote Sensing* **2020**, *12*, 2509.
28. Li, R.; Peng, Y.; Yang, Q. Fusion enhancement: UAV target detection based on multi-modal GAN. 2023 IEEE 7th Information Technology and Mechatronics Engineering Conference (ITOEC). IEEE, 2023, Vol. 7, pp. 1953–1957.
29. Liang, P.; Jiang, J.; Liu, X.; Ma, J. Fusion from decomposition: A self-supervised decomposition approach for image fusion. European Conference on Computer Vision. Springer, 2022, pp. 719–735.
30. Bao, Y.; Song, K.; Wang, J.; Huang, L.; Dong, H.; Yan, Y. Visible and thermal images fusion architecture for few-shot semantic segmentation. *Journal of Visual Communication and Image Representation* **2021**, *80*, 103306.
31. Dai, X.; Yuan, X.; Wei, X. TIRNet: Object detection in thermal infrared images for autonomous driving. *Applied Intelligence* **2021**, *51*, 1244–1261.
32. Wang, Q.; Chi, Y.; Shen, T.; Song, J.; Zhang, Z.; Zhu, Y. Improving RGB-infrared object detection by reducing cross-modality redundancy. *Remote Sensing* **2022**, *14*, 2020.
33. Luo, F.; Li, Y.; Zeng, G.; Peng, P.; Wang, G.; Li, Y. Thermal infrared image colorization for nighttime driving scenes with top-down guided attention. *IEEE Transactions on Intelligent Transportation Systems* **2022**, *23*, 15808–15823.
34. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
35. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11534–11542.
36. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794–7803.
37. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.
38. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. Proceedings of the IEEE/CVF international conference on computer vision workshops, 2019, pp. 0–0.
39. Li, X.; Hu, X.; Yang, J. Spatial group-wise enhance: Improving semantic feature learning in convolutional networks. *arXiv preprint arXiv:1905.09646* **2019**.
40. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 3146–3154.
41. Zhang, Q.L.; Yang, Y.B. Sa-net: Shuffle attention for deep convolutional neural networks. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 2235–2239.

 doi:10.20944/preprints202408.2054.v1

42. Wu, Y.; He, K. Group normalization. Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.
43. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.
44. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. Proceedings of the IEEE international conference on computer vision, 2017, pp. 764–773.
45. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.
46. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6154–6162.
47. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully convolutional one-stage object detection. arXiv 2019. *arXiv preprint arXiv:1904.01355* **2019**.
48. Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Single-shot refinement neural network for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4203–4212.
49. Li, S.; Li, Y.; Li, Y.; Li, M.; Xu, X. Yolo-firi: Improved yolov5 for infrared image object detection. *IEEE access* **2021**, *9*, 141861–141875.
50. Jiang, X.; Cai, W.; Yang, Z.; Xu, P.; Jiang, B. IARet: A lightweight multiscale infrared aerocraft recognition algorithm. *Arabian Journal for Science and Engineering* **2022**, *47*, 2289–2303.
51. Li, Q.; Zhang, C.; Hu, Q.; Fu, H.; Zhu, P. Confidence-aware fusion using dempster-shafer theory for multispectral pedestrian detection. *IEEE Transactions on Multimedia* **2022**, *25*, 3420–3431.
52. Chen, Q.; Wang, Y.; Yang, T.; Zhang, X.; Cheng, J.; Sun, J. You only look one-level feature. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 13039–13048.
53. Zheng, Z.; Wu, Y.; Han, X.; Shi, J. Forkgan: Seeing into the rainy night. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. Springer, 2020, pp. 155–170.
54. Anoosheh, A.; Sattler, T.; Timofte, R.; Pollefeys, M.; Van Gool, L. Night-to-day image translation for retrieval-based localization. 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019, pp. 5958–5964.
55. Liu, M.Y.; Breuel, T.; Kautz, J. Unsupervised image-to-image translation networks. *Advances in neural information processing systems* **2017**, *30*.