remote sensing

MDPI

Article

Contextual-Semantic Interactive Perception Network for Small Object Detection in UAV Aerial Images

Yiming Xu and Hongbing Ji *

Xi'an Key Laboratory of Intelligent Spectrum Sensing and Information Fusion, Xidian University, Xi'an 710071, China; xuyiming@stu.xidian.edu.cn
* Correspondence: hbji@xidian.edu.cn

Highlights

What are the main findings?

- An innovative Contextual-Semantic Interactive Perception Network (CSIPN) is proposed to address the challenges in small object detection, significantly improving small object detection performance in complex background through scene interaction modeling, dynamic context modeling, and semantic-context dynamic fusion.
- Three core modules are designed: the Scene Interaction Modeling Module (SIMM), the Dynamic Context Modeling Module (DCMM), and the Semantic Context Dynamic Fusion Module (SCDFM). These modules optimize the modeling of the relationship between small objects and the background, the adaptive modeling of contextual information, and the dynamic fusion of deep and shallow features, respectively.

What is the implication of the main finding?

- The Scene Interaction Modeling Module (SIMM) breaks through the limitations of traditional multi-scale feature fusion methods by explicitly modeling the implicit relationships between small objects and background under global semantic guidance. This significantly enhances the detection accuracy and robustness of small objects in complex background, providing a novel approach to small object detection.
- The Dynamic Context Modeling Module (DCMM) overcomes the limitations of fixed receptive field designs by flexibly adjusting receptive fields and dynamically generating weights. This allows the network to flexibly select contextual information for different small objects, effectively improving detection recall and preventing missed detections. This innovative design greatly enhances the model's adaptability and generalization in complex scenarios captured by unmanned aerial vehicles (UAVs).

Abstract

Unmanned Aerial Vehicle (UAV)-based aerial object detection has been widely applied in various fields, including logistics, public security, disaster response, and smart agriculture. However, numerous small objects in UAV aerial images are often overwhelmed by large-scale complex backgrounds, making their appearance difficult to distinguish and thereby prone to being missed by detectors. To tackle these issues, we propose a novel Contextual-Semantic Interactive Perception Network (CSIPN) for small object detection in UAV aerial scenarios, which enhances detection performance through scene interaction modeling, dynamic context modeling, and dynamic feature fusion. The core components of the CSIPN include the Scene Interaction Modeling Module (SIMM), the Dynamic Context Modeling Module (DCMM), and the Semantic-Context Dynamic Fusion Module (SCDFM). Specifically, the SIMM introduces a lightweight self-attention mechanism to generate a global scene semantic embedding vector, which then interacts with shallow spatial

check for updates

Academic Editor: Fa-Ming Fang

Received: 2 September 2025
Revised: 19 October 2025
Accepted: 28 October 2025
Published: 29 October 2025

Citation: Xu, Y.; Ji, H. Contextual-Semantic Interactive Perception Network for Small Object Detection in UAV Aerial Images. *Remote Sens.* **2025**, *17*, 3581. https://doi.org/10.3390/rs17213581

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

descriptors to explicitly depict the latent relationships between small objects and complex background, thereby selectively activating key spatial responses. The DCMM employs two dynamically adjustable receptive-field branches to adaptively model contextual cues and effectively supplement the contextual information required for detecting various small objects. The SCDFM utilizes a dual-weighting strategy to dynamically fuse deep semantic information with shallow contextual details, highlighting features relevant to small object detection while suppressing irrelevant background. Our method achieves mAPs of 37.2%, 93.4%, 50.8%, and 48.3% on the TinyPerson dataset, the WAID dataset, the VisDrone-DET dataset, and our self-built WildDrone dataset, respectively, while using only 25.3M parameters, surpassing existing state-of-the-art detectors and demonstrating its superiority and robustness.

Keywords: UAV; small object detection; scene interaction modeling; dynamic context modeling; dynamic feature fusion

1. Introduction

In recent years, the progressive maturation of precision sensors, advanced batteries, autonomous navigation systems, and computer vision technologies has enabled UAVs to achieve widespread adoption across a diverse range of application scenarios, including geospatial mapping [1], parcel delivery [2], public security patrols [3], emergency response [4], and agricultural monitoring [5]. Owing to their low deployment cost, broad observational coverage, and agile task scheduling, UAVs have demonstrated significant practical value in reducing manpower demands, accelerating operational workflows, and enhancing task safety. However, small object detection, which plays a vital role in advancing UAV intelligence, is still at a relatively immature stage [6,7]. This technical lag has led to unsatisfactory detection accuracy in complex UAV aerial scenarios, thereby posing a major constraint on the further adoption and evolution of UAVs in various industries. Although the advent of deep neural networks, particularly convolutional architectures, has markedly advanced object detection capabilities [8–11], existing algorithms still fail to meet the precision and robustness demands of UAV applications. The main challenges are as follows. First, UAV imagery often contains numerous small objects with subtle visual cues, making it difficult for conventional detection frameworks to extract sufficiently discriminative features, which in turn degrades detection accuracy. Second, the wide coverage and dynamic viewpoints inherent to aerial imagery introduce complex terrain textures and varying lighting conditions, both of which can significantly hinder the detection process and compromise the reliability of the results.

To address the above challenges, researchers have conducted extensive studies on enhancing the representation of small objects, among which multi-scale feature fusion has become a widely adopted approach for improving detection performance [12,13]. Although existing multi-scale fusion methods enhance small object detection by aggregating semantic and spatial information across different feature levels [14,15], they still suffer from several limitations. These methods typically rely on direct concatenation or a simple weighted summation of multi-scale feature maps. While such operations help alleviate the problem of insufficient feature expression, they lack explicit modeling of global semantic context, making it difficult to capture the implicit relationships between small objects and their surrounding environments. As a result, these methods are often inadequate in characterizing object–scene interactions, which compromises detection accuracy and robustness. In contrast, the proposed Scene Interaction Modeling Module (SIMM) introduces a simpli-

fied self-attention mechanism over deep features to generate compact scene embeddings, which are then combined with spatial descriptors obtained via the channel-wise pooling of shallow features. By performing interactive modeling between these two components, the SIMM explicitly incorporates scene-level priors into the fusion process. This enables the network to focus on the global contextual relationships between small objects and the scene, guiding the selective enhancement or suppression of shallow spatial responses under semantic supervision. Consequently, the SIMM significantly improves the representational distinctiveness and discriminative power of small object features, achieving more accurate localization and more reliable cross-scale alignment, particularly in complex background conditions.

Studies have shown that the contextual information surrounding small objects can provide valuable cues for object detection, and effectively leveraging this information helps improve detection performance [16,17]. Existing contextual information-based methods typically introduce fixed receptive fields to capture local or global context [18–20]. However, this fixed design has notable limitations: it fails to accommodate the varying contextual needs of small objects with different characteristics, and it may also introduce redundant background information that weakens the discriminability of object features. This issue is particularly pronounced in UAV aerial imagery, where small objects vary significantly in categories, textures, and fine-grained details. To address this problem, we propose the Dynamic Context Modeling Module (DCMM), which constructs two receptive field branches and dynamically generates their weights based on the input features. By selectively combining the responses of these branches in the spatial dimension, the DCMM enables adaptive context modeling. This design breaks through the limitations of fixed receptive fields, allowing the network to flexibly adjust the contextual information needed for different small objects. As a result, it facilitates feature enhancement in an object-specific manner and effectively mitigates the problem of missed detections caused by the improper use of contextual information.

To comprehensively enhance the representation of small objects from the perspectives of scene understanding, context modeling, and semantic fusion, we propose a novel detector tailored for UAV aerial imagery, which is named the Context–Semantic Interaction Perception Network (CSIPN). This network comprises three key modules: the SIMM, the DCMM, and the Semantic-Context Dynamic Fusion Module (SCDFM). Specifically, in the scene interaction modeling stage, the SIMM constructs a simplified self-attention mechanism on deep features to generate compact scene embedding vectors while applying channel-wise pooling operations to shallow features to obtain spatial descriptors. These two representations are then jointly modeled to explicitly capture the semantic correlations between small objects and the global scene, enabling the selective enhancement or suppression of shallow spatial responses under the guidance of global semantic priors. This process improves the discriminability and saliency of small object representations. In the context modeling stage, the DCMM builds context modeling branches with diverse receptive fields and dynamically assigns weights based on input features, enabling the selective integration and enhancement of contextual information. By overcoming the limitations of fixed receptive fields, the DCMM strengthens the network's ability to adaptively capture the contextual cues needed by different small objects, thereby improving detection recall. In the semantic-context dynamic fusion stage, the SCDFM adaptively integrates the deep semantic features extracted from the backbone with the context-enhanced representations generated by the DCMM, supplementing and balancing semantic and contextual cues to facilitate comprehensive feature interactions and refine small object representations. Most existing UAV aerial datasets primarily focus on urban traffic scenarios, leading to deep learning-based methods that often suffer from poor generalization and limited robustness

in real-world applications. To address this issue, we construct a new UAV-based object detection dataset named WildDrone, objecting diverse and unstructured wild environments, with the aim of enhancing model adaptability and practical performance in complex scenes. Comprehensive evaluations are performed on the VisDrone-DET [21], TinyPerson [22], WAID [23], and WildDrone datasets. Experimental results show that the CSIPN surpasses current state-of-the-art detectors in both detection accuracy and robustness.

1. A novel Context–Semantic Interaction Perception Network (CSIPN) is proposed, which integrates scene understanding, dynamic context modeling, and semantic fusion to effectively improve the detection performance of small objects in UAV aerial imagery.
2. A Scene Interaction Modeling Module (SIMM) is designed to explicitly capture the relationships between small objects and the global scene. Shallow spatial responses are guided to be selectively enhanced or suppressed under global semantic priors, thereby strengthening semantic perception and discriminative modeling in complex background.
3. A Dynamic Context modeling Module (DCMM) is presented, in which multi-receptive-field branches are employed with dynamically assigned weights. This enables the adaptive selection and integration of contextual information across different scales, facilitating the dynamic compensation of missing appearance cues for small objects and improving detection completeness and robustness.
4. A Semantic-Context Dynamic Fusion Module (SCDFM) is proposed, which adaptively integrates the deepest semantic information with the shallow contextual information, effectively leveraging their complementary relationship to enhance the representations of small objects.
5. To improve the generalization and practical applicability of detection models in complex environments, a new UAV-based dataset named WildDrone is constructed, which focuses on diverse and unstructured wild scenes.
6. Extensive experiments on UAV aerial datasets including VisDrone-DET, TinyPerson, WAID, and WildDrone validate the superior performance of the proposed method in small object detection tasks.

2. Related Work

Strategies for enhancing small object detection performance can be generally divided into three categories according to their primary enhancement mechanisms: multi-scale representation-based, context-based, and super-resolution (SR)-based methods. This categorization is consistent with previous studies [24–27] and reflects the main directions that current approaches take to address the challenges of small-scale objects.

2.1. Multi-Scale Representation-Based Methods

Multi-scale representation-based methods allow the network to leverage features at varying spatial resolutions. In these feature maps, deep features with low resolution are rich in semantic information, aiding in the classification of small objects, while shallow features with high resolution preserve more detailed spatial information, enhancing the accurate localization of small objects [28]. Lin et al. proposed the Feature Pyramid Network (FPN) [29], which merges low- and high-level features through a top–down connection path, thereby helping detect objects of different scales. Min et al. proposed an attentional FPN (AFPN) [30], which enhances small object detection capabilities by designing dynamic texture attention, foreground-aware co-attention, and detail context attention. Ge et al. introduced an adaptive reparameterized generalized FPN (Adaptive-RepGFPN) [31], which enhances the feature representation of small objects by reorganizing features and incorpo-

rating adaptive weighting in the concatenation operation. HR-FPN [26] boosts the accuracy of small object detection through the stepwise alignment and fusion of high-resolution features, all while preserving the network's lightweight design. Kiobya et al. presented a multi-scale semantic enhanced FPN (MSSEFPN) [32], which extracts important detailed features from high-level prediction layers and adaptively fuses them into the lowest-level prediction layers, significantly enhancing the semantic representation of small objects.

Although these methods fully utilize multi-scale features to enrich semantic information, they neglect the thorough exploration of contextual information. For small objects with low resolution and missing detail information, the lack of necessary prior knowledge often leads to missed detections or misclassifications. In addition to multi-scale feature enhancement, recent studies have also begun to explore cross-modal feature fusion strategies to further improve detection robustness in complex scenarios. For example, Hu et al. proposed a complementarity-aware feature fusion detection network (CFFDNet) [33], which combines differential feature spatial-aware complementary units and gate-generated weighted fusion units to effectively enhance multimodal feature integration, providing a potential direction for improving feature representation in small object detection tasks.

2.2. Context-Based Methods

2.2.1. Conventional Methods

Context-based methods significantly improve the accuracy of small object detection by deeply exploring and effectively utilizing global or local context. The feature fusion and scaling-based single shot detector (FSSSD) [34] introduces spatial context analysis, which effectively integrates the spatial relationships between objects during the object re-identification process, optimizing small object detection in UAV aerial imagery. CAB Net [35] uses dilated convolutions to fuse multi-level contextual information, generating high-resolution feature maps rich in semantics, thereby further enhancing the detection capability of small objects. Shi et al. optimized multi-level feature expressions by merging contextual and semantic information [36], improving the localization and recognition of small objects in remote sensing images. Chen et al. proposed a hybrid receptive field network [18] that extends shallow features with different receptive fields and captures cross-layer contextual information, significantly improving the ability to perceive small objects. Yang et al. designed pinwheel-shaped convolution (PConv) [20], which adapts to the spatial distribution characteristics of infrared small objects, significantly expanding the receptive field with minimal additional parameters, thereby enhancing context modeling and improving the detection accuracy of small objects.

2.2.2. Self-Attention-Based Methods

Self-attention mechanisms can naturally capture long-range dependencies within feature maps, enabling the network to establish relationships between each spatial location and all others. This global modeling capability effectively encodes contextual relationships, thereby enhancing the network's discriminative power for small objects. Consequently, self-attention-based methods inherently belong to the context-based detection paradigm, as they explicitly leverage contextual information to guide the detection process. ACmix [19] combines self-attention mechanisms with convolutional operations to effectively harness both global and local context, thus enhancing the feature representation of small objects. Hu et al. [37] proposed CM-YOLO, a remote sensing object detection framework for cloud and mist scenarios, which employs hierarchical selective attention alongside convolutional feature extraction to jointly mine global and local semantic context, effectively enhancing target features. MAFDet [38] introduces a multi-attention focusing sub-network that generates attention maps for the precise localization of densely distributed small objects,

and a multi-scale Swin Transformer backbone that enhances multi-layer feature extraction while effectively suppressing background interference. STDet [39] constructs a local-global Transformer to establish associations between tokens and surrounding data while capturing long-range dependencies related to the objects. It further designs a scale-balanced label assignment strategy to dynamically adjust the learning focus toward easily overlooked small objects, thereby alleviating sample imbalance and enhancing small object detection.

In summary, although existing context-based methods significantly enhance small object detection performance, they still face inherent limitations. Due to the complexity of UAV aerial scenes, the variability of object types and scales, and the different contextual information required by different objects, the fixed receptive field limits the effective establishment of appropriate context relationships, making the effective utilization of contextual information difficult.

2.3. SR-Based Methods

SR-based methods enhance low-resolution images or regions of interest (RoIs) by converting them into high-resolution ones, thereby providing the network with more detailed feature representations of small objects. Hu et al. [40] successfully improved small object localization accuracy by applying bilinear interpolation to large input images. However, performing SR on the entire image is inefficient because background regions, which are unrelated to the detection task, are also processed, leading to a waste of computational resources. To address this issue, SOD-MTGAN [41] first uses a high-recall detector to extract RoIs and then performs SR processing only on these regions, significantly improving computational efficiency. Rabbi et al. proposed an edge-enhanced SR generative adversarial network (EESRGAN) [42], which combines ESRGAN, edge-enhancement network (EEN), and detection networks to improve remote sensing image quality and enhance small object detection performance. Bashir et al. proposed a detection framework based on residual feature aggregation-based SR [43], which integrates cyclic GANs and residual feature aggregation (RFA) technology to improve small object detection in satellite or aerial remote sensing images. Zhang et al. [44] enhanced the feature representation of small objects by introducing an auxiliary SR learning mechanism, allowing the network to maintain excellent detection performance even under low-resolution input conditions. Liu et al. [45] jointly trained the SR task with the object detection task to guide the backbone network to learn high-resolution features, significantly improving small object detection accuracy. Although SR-based methods effectively supplement the detailed information needed for small object detection, these methods still have limitations: during reconstruction, unreal features or artifacts may be introduced, causing the reconstructed region's features to deviate from the real scene, which increases background noise interference and leads to false detections.

3. Methods

3.1. Overview of CSIPN

The overall architecture of the CSIPN is illustrated in Figure 1. The network consists of five stages: backbone feature extraction, scene interaction modeling, dynamic context modeling, semantic-context fusion, and prediction. In the backbone feature extraction stage, the network extracts fundamental features from the input image, generating multi-level feature representations for subsequent modeling. These features are denoted as $P_i \in \mathbb{R}^{C_i \times H_i \times W_i}, i \in 1, 2, 3, 4$, where C_i, H_i, and W_i represent the number of channels, height, and width of the feature map at level i, respectively.

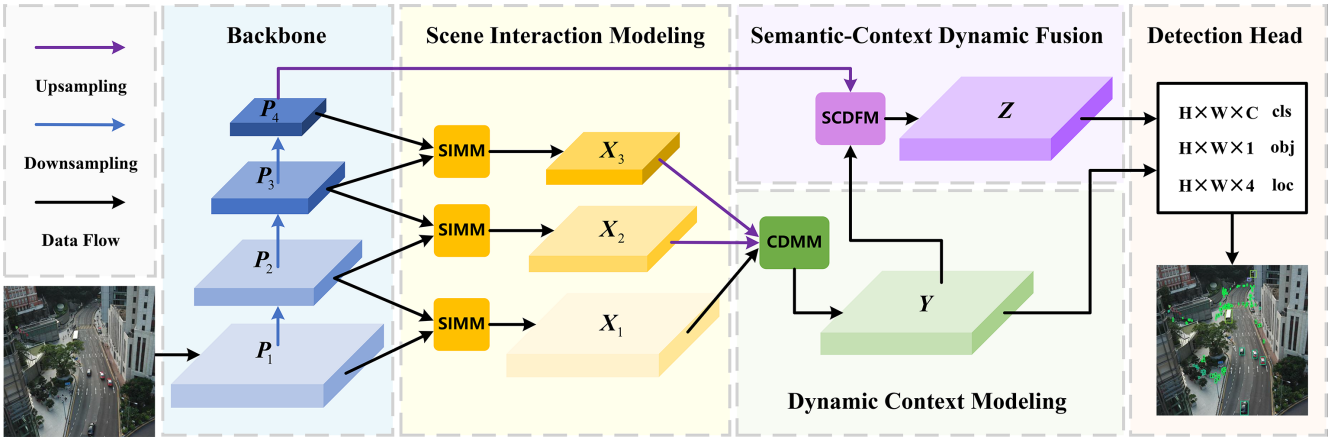

Figure 1. Overall framework of SFFPNet.

In the scene interaction modeling stage, the network employs the SIMM to establish the relationships between small objects and the global scene. This module leverages deep semantic cues to guide the spatial responses of shallow features, thereby achieving implicit semantic alignment and providing semantic priors for subsequent context modeling.

In the dynamic context modeling stage, the DCMM dynamically aggregates contextual information from multiple receptive field branches under the guidance of the semantic priors provided by the SIMM. Through multi-branch weighting and spatially selective fusion, the DCMM adaptively captures contextual dependencies suited to different objects, effectively enhancing the discriminative power of foreground features and alleviating missed detections in UAV aerial imagery.

In the semantic-context fusion stage, the SCDFM dynamically fuses the deep semantic features extracted by the backbone with the context-enhanced representations generated by the DCMM. This module adaptively selects and integrates complementary information from different features, ensuring that the fused representations are semantically consistent and mutually complementary.

Finally, in the prediction stage, the features produced by the DCMM and SCDFM are fed into the detection heads to generate the final detection results. Overall, the CSIPN establishes a progressive and collaborative feature modeling framework: the SIMM provides semantic guidance while achieving cross-scale fusion; under this guidance, the DCMM performs dynamic context modeling; and the SCDFM further builds a dynamic complementarity between semantics and context, thereby enhancing the overall discriminability and robustness for small object detection.

3.2. Scene Interaction Modeling Module (SIMM)

Different from conventional Transformer-based global attention mechanisms that rely on full self-attention computation, the SIMM introduces a lightweight and semantically guided interaction strategy that efficiently captures global contextual dependencies while selectively enhancing shallow spatial details, thereby improving the saliency and discriminability of small object representations. Specifically, the SIMM approximates global attention through a series of 1×1 convolutional and reshaping operations to generate a compact global scene embedding vector without computing the full $N \times N$ attention matrix, where N denotes the total number of pixels in the input feature map. This design significantly reduces computational complexity while maintaining the capability to capture long-range dependencies. More importantly, by integrating this global semantic prior with shallow spatial responses, the SIMM explicitly models the interaction between global semantics and fine-grained spatial information, which distinguishes it from existing global attention modules.

To further mitigate the semantic inconsistency between deep and shallow features during multi-scale fusion, the SIMM incorporates a guided interaction mechanism to achieve implicit semantic alignment. The scene embedding E_i generated from deep features acts as a semantic prior to modulate the spatial responses D_i of shallow features. This semantic guidance selectively enhances the spatial details associated with small objects while suppressing irrelevant background interference. As a result, the shallow spatial responses are filtered and refined under the guidance of global semantics, achieving alignment with the global scene information extracted from deep features. Consequently, the fused features across different scales preserve both spatial coherence and semantic consistency, leading to more robust and discriminative multi-scale representations that enhance small object detection performance. The detailed structure of this module is shown in Figure 2.

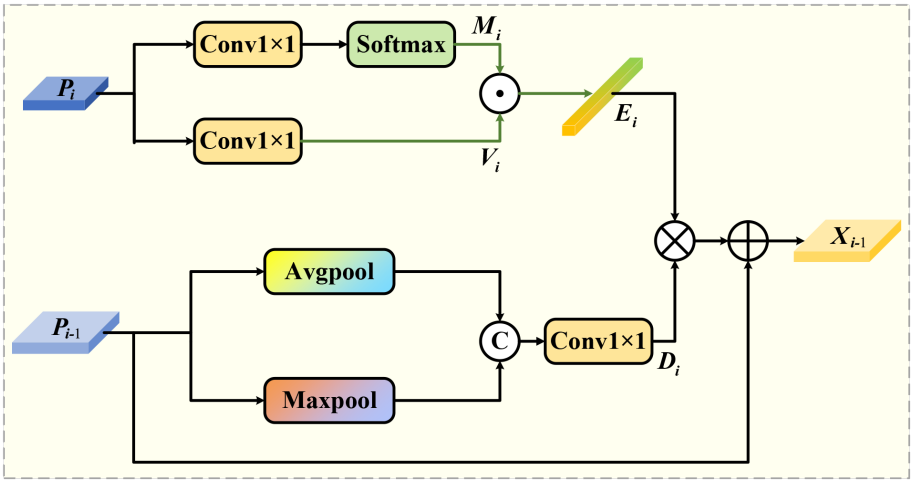

Figure 2. Structure of SIMM.

We first design a lightweight self-attention mechanism to perform global relation modeling on the deep feature P_i that contains richer semantic information, aiming to capture long-range dependencies between pixels. Specifically, we approximate the similarity between the Query and Key matrices through convolution operations and feature dimension reshaping, which significantly reduces computational overhead while maintaining effective modeling capability. The detailed process is as follows:

$$M_i = \mathcal{R}(\mathcal{S}(\mathcal{C}_{1\times1}(P_i))) \tag{1}$$

where $M_i \in \mathbb{R}^{1\times H_i W_i \times 1}$ denotes the similarity matrix, $\mathcal{R}(\cdot)$ denotes the reshape operation, $\mathcal{S}(\cdot)$ denotes the Softmax function, and $\mathcal{C}_{n\times n}$ denotes an $n \times n$ convolution layer with a SiLU activation function. The value matrix $V_i \in \mathbb{R}^{1\times C_{i-1}\times H_i W_i}$ is generated as follows:

$$V_i = \mathcal{R}(\mathcal{C}_{1\times1}(X_i)) \tag{2}$$

Subsequently, we integrate M_i and V_i to construct a one-dimensional scene embedding representation $E_i \in \mathbb{R}^{C_{i-1}\times1\times1}$, which captures the semantic correlation between foreground small objects and the scene, and it is expressed as follows:

$$E_i = M_i \odot V_i \tag{3}$$

where $\odot$ denotes the function used to estimate the similarity between features. To accelerate computation, we efficiently implement this process using matrix multiplication.

To comprehensively capture the spatial fine-grained information of small objects, channel-wise max pooling and channel-wise average pooling are applied to the shallow feature P_{i-1}, followed by a convolution layer that fuses these two types of spatial descriptors to obtain the spatial detail response $D_i \in \mathbb{R}^{1\times H_{i-1}\times W_{i-1}}$:

$$D_i = \mathcal{C}_{1\times 1}[\mathrm{Avg}(P_{i-1}); \mathrm{Max}(P_{i-1})] \tag{4}$$

where $[;]$ denotes channel-wise concatenation, $\mathrm{Avg}(\cdot)$ represents channel-wise average pooling, and $\mathrm{Max}(\cdot)$ indicates channel-wise max pooling.

To selectively enhance or suppress shallow spatial responses under the guidance of global semantic priors and thereby improve the saliency and discriminability of small objects, the one-dimensional scene embedding E_i is applied to weight the spatial detail response D_i. A residual structure is then applied to ensure training stability, resulting in the final scene interaction feature $X_{i-1} \in \mathbb{R}^{C_{i-1}\times H_{i-1}\times W_{i-1}}$, which is expressed as

$$X_{i-1} = P_{i-1} \oplus (E_i \otimes D_i) \tag{5}$$

where $\oplus$ denotes element-wise addition and $\otimes$ denotes broadcast element-wise multiplication.

3.3. Dynamic Context Modeling Module (DCMM)

Due to the lack of distinctive appearance features in small objects, utilizing contextual information can be beneficial for detection to a certain extent. However, the excessive supplementation of contextual information may introduce a significant amount of irrelevant background and noise, reducing the detector's discriminative ability and increasing the false detection rate. Considering the diversity of object types and scales in UAV aerial images, the amount of contextual information required for detecting different objects also varies.

To address this, the DCMM is designed. It consists of two receptive field branches, whose weights (i.e., spatial selection masks) are dynamically determined based on the input. It is worth noting that the DCMM does not require any additional loss function to guide the learning of dynamic weights. Specifically, the dynamic weights depend both on the convolutional parameters learned during training and the characteristics of the input feature maps. The learned convolutional parameters define the rules for weight generation, while the input feature maps provide the information necessary to instantiate these rules. During training, the gradients from the detection loss are backpropagated through the DCMM, automatically optimizing the convolutional parameters that generate the spatial selection masks. During inference, these learned convolutional parameters interact with the input feature maps to dynamically generate the weights of the receptive field branches.

By selectively combining these receptive field branches in the spatial dimension, the DCMM overcomes the limitation of fixed receptive fields and enables the dynamic selection of contextual information tailored to different small objects. This allows the DCMM to adaptively supplement the contextual cues needed for detecting various small objects, enhancing distinct foreground features and effectively alleviating the issue of severe missed detections in UAV aerial imagery.

As shown in Figure 3, the input to the DCMM consists of X_1, X_2, and X_3. First, X_1 is upsampled by a factor of 4 and X_2 is upsampled by a factor of 2. Then, the upsampled X_1, upsampled X_2, and X_3 are concatenated along the channel dimension and passed through a 1×1 convolution layer to adjust the number of channels, resulting in $X_c \in \mathbb{R}^{C\times H_1\times W_1}$:

$$X_c = [X_1 \uparrow_{4\times}; X_2 \uparrow_{2\times}; X_3] \tag{6}$$

where $\uparrow_{nx}$ denotes n-times upsampling using the nearest-neighbor interpolation method.

Then, the fused feature X_c passes through two receptive field branches, producing outputs $R_1 \in \mathbb{R}^{C\times H_1\times W_1}$ and $R_2 \in \mathbb{R}^{C\times H_1\times W_1}$:

$$\begin{aligned} R_1 &= \mathcal{C}_{3\times3}(X_c) \oplus \mathcal{C}_{3\times3}(\mathcal{C}_{1\times1}(\mathcal{C}_{3\times3}(X_c))) \\ R_2 &= \mathcal{C}_{1\times1}(X_c) \oplus \mathcal{C}_{1\times1}(\mathcal{C}_{1\times1}(\mathcal{C}_{3\times3}(X_c))) \end{aligned} \tag{7}$$

The outputs from the different receptive field ranches R_1 and R_2 are concatenated to form $R_c \in \mathbb{R}^{2C\times H_1\times W_1}$:

$$R_c = [R_1; R_2] \tag{8}$$

Next, channel-wise max pooling and channel-wise average pooling are applied to R_c to extract the spatial relationship:

$$\begin{aligned} R_{avg} &= \mathrm{Avg}(R_c) \\ R_{max} &= \mathrm{Max}(R_c) \end{aligned} \tag{9}$$

where $R_{avg} \in \mathbb{R}^{1\times H_1\times W_1}$ and $R_{max} \in \mathbb{R}^{1\times H_1\times W_1}$ denote spatial feature descriptors obtained through max pooling and average pooling, respectively. To achieve information interaction between different spatial feature descriptors, R_{avg} and R_{max} are concatenated and passed through a 1×1 convolution layer with a SiLU activation function, producing the spatial selection mask $S \in \mathbb{R}^{2\times H_1\times W_1}$:

$$S = [S_1; S_2] = \mathcal{C}_{1\times1}([R_{avg}; R_{max}]) \tag{10}$$

The features from different receptive field branches are then weighted by their corresponding spatial selection masks and fused through a 1×1 convolution layer with a SiLU activation function to obtain the attention feature $R_a \in \mathbb{R}^{C\times H_1\times W_1}$:

$$R_a = \mathcal{C}_{1\times1}((R_1 \otimes S_1) \oplus (R_2 \otimes S_2)) \tag{11}$$

Finally, the output $Y \in \mathbb{R}^{C\times H_1\times W_1}$ of the DCMM is the element-wise product of X_c and R_a, which is expressed as shown below:

$$Y = X_c \otimes R_a \tag{12}$$

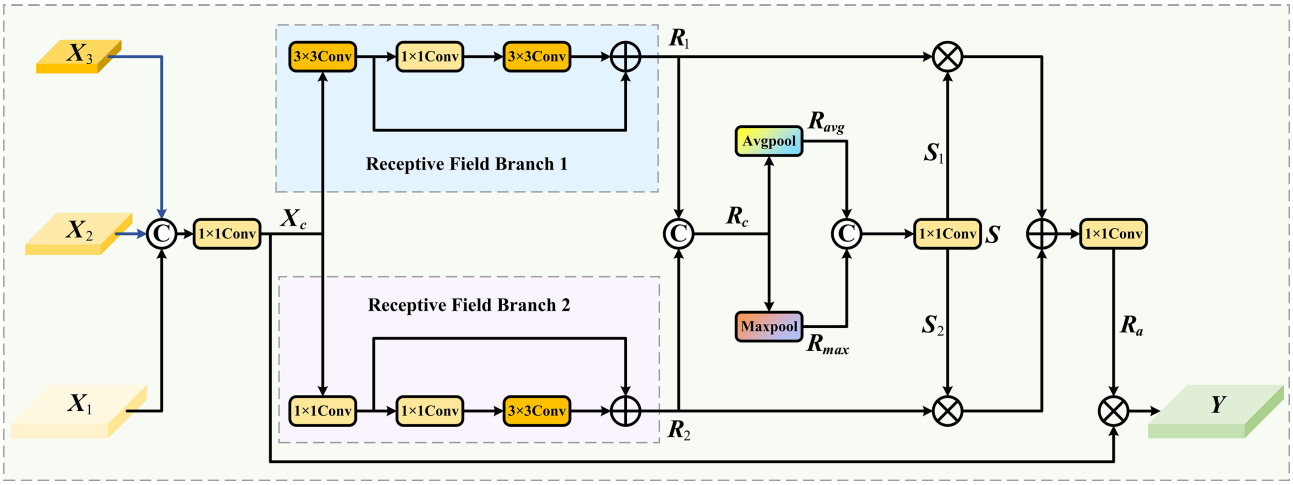

Figure 3. Structure of DCMM.

3.4. Semantic-Context Dynamic Fusion Module (SCDFM)

Unlike the SIMM and DCMM, which focus on semantic-guided scene interaction and contextual modeling, respectively, the SCDFM serves as a complementary fusion module that dynamically supplements semantic cues beneficial for small-object classification. Rather than performing explicit geometric alignment, the SCDFM concentrates on ensuring semantic compatibility between the modulated contextual information from the DCMM and the deep semantic representations extracted from the backbone. To this end, the SCDFM adaptively adjusts the fusion ratios of multi-scale features through channel-wise and spatial-wise attention, which is guided by deep semantics. This design allows the network to reconcile high-level semantic information with shallow contextual cues, thereby enhancing the discriminability of small-object features.

As shown in Figure 4, to ensure dimensional compatibility with shallow features, P_4 is upsampled by a factor of 8 and then passed through a 1×1 convolution layer to obtain $H \in \mathbb{R}^{C_4 \times H_1 \times W_1}$:

$$H = \mathcal{C}_{1\times1}(P_4 \uparrow_{8\times}) \tag{13}$$

To effectively extract key semantic information from deep features, global average pooling is applied to capture the overall distribution characteristics. The resulting tensor is then reshaped, which is followed by a one-dimensional convolution to perform nonlinear feature mapping. Finally, a Sigmoid activation function is used to generate the global channel descriptor $G_c \in \mathbb{R}^{C_4 \times 1 \times 1}$:

$$G_c = \text{Sigmoid}\Big(\mathcal{C}_3\Big(\mathcal{R}\Big(\mathcal{A}_{\text{global}}\,(P_4)\Big)\Big)\Big) \tag{14}$$

where $\text{Sigmoid}(\cdot)$ denotes the Sigmoid activation function, $\mathcal{C}_3$ represents a one-dimensional convolution layer with a kernel size of 3, and $\mathcal{A}_{\text{global}}$ indicates global average pooling. Then, G_c guides the selection of semantic information by performing a broadcast element-wise multiplication with H, resulting in the attention-enhanced feature $A \in \mathbb{R}^{C_4 \times H_1 \times W_1}$:

$$A = G_c \otimes H \tag{15}$$

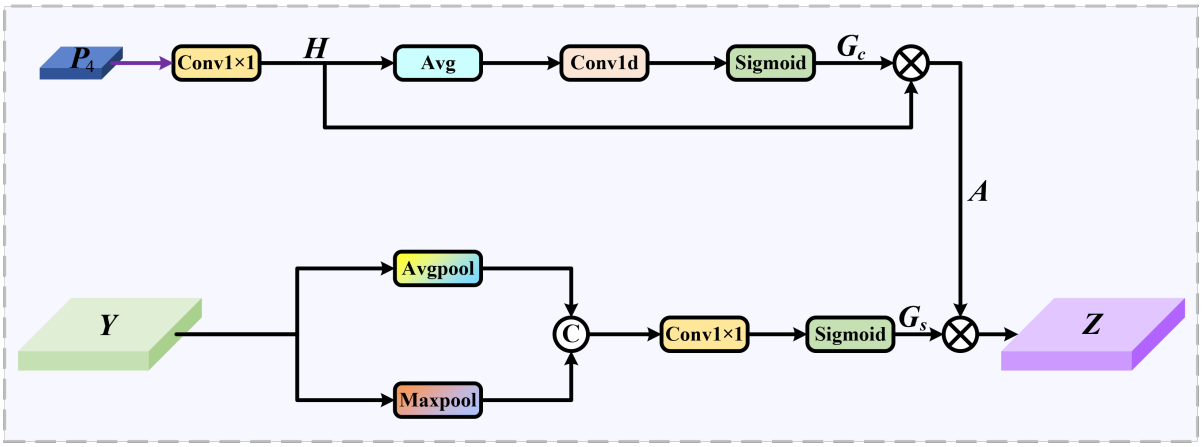

Figure 4. Structure of SCDFM.

To establish a feature selection mechanism along the spatial dimension, it is necessary to model the spatial contextual relationships of Y. Specifically, average pooling and max pooling are applied along the channel dimension to capture different statistical characteristics. The resulting feature is then transformed by a 1×1 convolution layer and passed through a Sigmoid function to generate the global spatial attention map $G_s \in \mathbb{R}^{1 \times H_1 \times W_1}$:

$$G_s = \text{Sigmoid}(\mathcal{C}_{1\times1}([\text{Avg}(Y); \text{Max}(Y)])) \tag{16}$$

Finally, the spatial descriptor G_s is used to guide spatially selective feature fusion, thereby highlighting key regions of small objects and producing the optimized output feature $Z \in \mathbb{R}^{C_4 \times H_1 \times W_1}$:

$$Z = G_s \otimes A \tag{17}$$

4. Results

This section begins with a comprehensive overview of the datasets and experimental setup. Subsequently, we evaluate the performance of each module through a series of ablation experiments. Finally, we test the performance of the CSIPN on multiple UAV aerial datasets and compare it with state-of-the-art methods.

4.1. Datasets

To highlight the advantages of our method in small object detection, we selected the VisDrone-DET dataset, the TinyPerson dataset, and our self-constructed WildDrone UAV aerial dataset for experimentation. Table 1 shows the distribution of different object sizes in these UAV aerial datasets compared to common datasets like Pascal VOC [46] and MS COCO [47]. The results indicate that the number of small objects in UAV aerial images is significantly higher than in traditional datasets, which undoubtedly presents a greater challenge for detectors.

Table 1. Comparison of the proportion of objects at different scales across multiple datasets. The images in WAID are resized to 448 × 448, while those in other datasets are uniformly resized to 896 × 896.

Datasets	Type	Small (%)	Medium (%)	Large (%)
Pascal VOC	General	1.58	14.76	83.66
MS COCO	General	15.79	33.22	50.98
VisDrone-DET	UAV	74.36	23.72	1.92
TinyPerson	UAV	97.05	2.86	0.09
WAID	UAV	76.61	22.75	0.64
WildDrone	UAV	75.30	21.26	3.44

4.1.1. VisDrone-DET

This dataset is specifically designed for UAV aerial imagery of urban scenes. The training set consists of 6471 images, the test set contains 1610 images, and the validation set includes 548 images, with a total of approximately 457,100 annotated bounding boxes. The objects in these images are categorized into 10 classes, including pedestrian, people, bicycle, car, van, truck, tricycle, awning-tricycle, bus, and motor.

4.1.2. TinyPerson

This dataset provides a benchmark for detecting tiny objects in long-range shots and complex backgrounds, consisting of 1610 UAV-captured images, with 794 images in the training set and 816 in the test set. The TinyPerson dataset contains 72,651 small person objects, which are each manually annotated and categorized into five classes: Ocean, Earth, Ignore, Uncertain, and Dense. During the experiments, we unified all annotated small person objects into the "Person" category for processing.

4.1.3. WAID

This is a large-scale UAV aerial dataset focused on wildlife detection. The training, validation, and test sets contain 10,056, 2873, and 1437 UAV aerial images, respectively, with 233,806 object bounding boxes in total. The dataset includes six wildlife categories: sheep, cattle, seal, camelus, kiang, and zebra. WAID covers diverse environmental con-

ditions and multiple habitat types, making it suitable for developing and evaluating UAV-based wildlife detection algorithms.

4.1.4. WildDrone

Currently, the vast majority of UAV aerial datasets focus on relatively simple scenes (mostly urban environments), where the background surrounding small objects is usually limited to roads. This leads to deep learning-based detectors exhibiting poor robustness and generalization ability when faced with complex and dynamic backgrounds, thus limiting their range of applications. To overcome this limitation, we have developed a UAV perspective wild scene object detection dataset, named WildDrone, which covers a variety of natural terrains, including forests, plains, mountains, canyons, oceans, and snowfields. Small objects are often hidden in these complex backgrounds. The dataset contains 1950 images, with 1260 for training and 690 for testing. The WildDrone dataset includes approximately 23,567 annotated object instances, which are categorized into "person" and "vehicle" classes. Figure 5 shows some typical scenes from this dataset, highlighting its rich diversity and challenges.



Figure 5. Visualization of diverse scenarios in WildDrone.

4.2. Implementation Details

All training and testing of the models are conducted on a server equipped with a 24-core CPU, 64GB of RAM, and an RTX 4090 GPU. During training, we use the SGD optimization algorithm with a batch size of 4. Due to memory limitations, the batch size is reduced to 2 when training DINO [48]. The initial learning rate is set to 0.01, and dynamic adjustment is performed using cosine decay. We choose the convenient YOLOv11 [49] as the baseline. Since our method differs significantly from the baseline in structure, no pretrained weights are used. Our method follows the same loss design as the baseline, which includes classification, localization, and confidence losses. Specifically, we adopt the standard loss formulation used in YOLOv11 to ensure a fair comparison and a consistent optimization objective. During testing, the confidence threshold is set to 0.15, and the IoU threshold is set to 0.6. In the experiments on the WAID dataset, the resolution of the input

images is set to 448 × 448, while in the experiments on other datasets, it is uniformly set to 896 × 896.

4.3. Evaluation Metrics

To evaluate the detection accuracy, we select average precision (AP) and mean average precision (mAP) as the evaluation metrics. For the single-class TinyPerson dataset, we use the AP metric for performance validation. For multi-class scenarios like the VisDrone2019-DET and WildDrone datasets, we use mAP for evaluation. Although the MS COCO [47] protocol typically uses AP values for areas smaller than 32^2 to reflect small object detection performance, this approach proves insufficient when handling UAV aerial datasets with a large number of tiny objects. Therefore, we introduce a more detailed metric system, which includes the following: APs1 (area < 20^2), APs2 (20^2 < area < 24^2), APs3 (24^2 < area < 28^2), APs4 (28^2 < area < 32^2), APs (area < 32^2), and AP values for all objects. In the experiments on the VisDrone-DET and WildDrone datasets, the IoU threshold for mAP calculation is fixed at 0.5. For the TinyPerson dataset, the IoU threshold for AP calculation is set to 0.25. To avoid truncation in dense scenes, the maximum number of detections per image is limited to 1000.

4.4. Ablation Studies

To evaluate the detection performance of each module in CSIPN, we selected YOLOv11-L as the baseline and conducted ablation experiments on the single-category TinyPerson test set and the multi-category WAID test set. Tables 2 and 3 present the results of the ablation study.

Table 2. Results of the ablation studies on the TinyPerson test set. Bold font indicates the best result of one metric.

Method	SIMM	DCMM	SCDFM	APs1	APs2	APs3	APs4	APs	AP
Baseline				19.0	40.4	45.6	52.1	24.0	28.1
	✓			27.9	48.2	51.9	55.6	31.3	34.7
	✓	✓		30.5	52.7	56.6	57.9	35.0	36.9
	✓	✓	✓	**32.6**	**55.1**	**58.4**	**59.9**	**37.2**	**41.2**

4.4.1. Effectiveness of SIMM

We introduce the SIMM into the backbone to achieve scene interaction modeling. As shown in rows 1 and 2 of Table 2, with the assistance of this module, the baseline achieves improvements of 8.9%, 7.8%, 6.3%, 3.5%, and 7.3% on APs1, APs2, APs3, APs4, and APs, respectively. Similarly, as shown in rows 1 and 2 of Table 3, the SIMM helps the baseline achieve improvements of 4.8%, 3.5%, 2.2%, 1.8%, and 3.9% on mAPs1, mAPs2, mAPs3, mAPs4, and mAPs, respectively, fully demonstrating the remarkable effectiveness of the SIMM in small object detection. In addition, the overall AP and mAP increase by 6.6% and 1.0% in Tables 2 and 3, respectively, further validating its effectiveness in complex UAV aerial scenes. These results confirm that the SIMM can explicitly establish the relationship between small objects and the global scene, enhancing the shallow feature responses to key regions under semantic guidance, thereby significantly improving the detection accuracy for small objects.

Moreover, as the first module in the pipeline, the SIMM provides semantic priors that guide subsequent modules (DCMM and SCDFM). This highlights its foundational role in establishing meaningful scene interactions, which enables the later modules to exploit context and perform adaptive fusion more effectively, thus contributing to the overall synergistic improvement of the CSIPN.

Table 3. Results of the ablation studies on the WAID test set. Bold font indicates the best result of one metric.

Method	SIMM	DCMM	SCDFM	mAPs1	mAPs2	mAPs3	mAPs4	mAPs	mAP
Baseline				68.4	83.2	90.9	92.6	86.4	94.6
	✓			73.2	86.7	93.1	94.4	90.3	95.6
	✓	✓		75.8	89.0	94.2	95.8	92.5	96.3
	✓	✓	✓	**77.5**	**90.2**	**94.9**	**96.7**	**93.4**	**96.7**

4.4.2. Effectiveness of DCMM

We further introduce the DCMM into the network to enable context modeling. As shown in rows 2 and 3 of Table 2, with the aid of this module, the network achieves improvements of 2.6%, 4.5%, 4.7%, 2.3%, and 3.7% on APs1, APs2, APs3, APs4, and APs, respectively. Similarly, as shown in rows 2 and 3 of Table 3, with the assistance of this module, the network achieves improvements of 2.6%, 2.3%, 1.1%, 1.4%, and 2.2% on mAPs1, mAPs2, mAPs3, mAPs4, and mAPs, respectively. This indicates that the DCMM effectively enhances small object detection performance by adaptively integrating contextual information from different receptive fields to meet the varying perception needs of small objects across diverse backgrounds. Moreover, the DCMM further increases the overall AP and mAP by 2.2% and 0.7%, respectively, demonstrating its strong adaptability to diverse small-object scenarios and its contribution to enhancing the overall detection capability.

Importantly, the DCMM operates on the semantic priors generated by the SIMM, allowing it to dynamically focus on foreground-relevant context guided by scene semantics. This dependency illustrates a collaborative effect, where the context modeling performed by the DCMM is more effective when the preceding semantic guidance from the SIMM is available, reflecting the nonlinear synergistic interaction between these modules.

To investigate the effect of different input feature combinations in the DCMM, we conducted ablation studies under four schemes. Table 4 presents the detection performance of each scheme. As shown, Scheme 1 achieves the best mAPs, demonstrating that integrating multi-scale features into a single DCMM effectively captures complementary contextual information. Scheme 2 slightly outperforms Scheme 3 because combining low- and mid-level features provides richer spatial details critical for small object detection, whereas Scheme 3 relies more on mid- and high-level features, resulting in coarser spatial granularity. Scheme 4 exhibits the lowest performance due to fragmented context modeling, in which each DCMM independently processes only a single-scale feature map, resulting in incomplete contextual representation and suboptimal dynamic weight learning. These results validate our design choice illustrated in Figure 3, where X_1, X_2, and X_3 are fed into a single DCMM.

Table 4. Detection performance of DCMM with different input feature combinations. Bold font indicates the best result for each metric.

Scheme	Methods	mAPs1	mAPs2	mAPs3	mAPs4	mAPs
1	X_1, X_2, X_3 → one DCMM	**30.5**	**52.7**	**56.6**	**57.9**	**35.0**
2	X_1, X_2 → DCMM_1; X_3 → DCMM_2	30.1	52.2	56.1	57.5	34.5
3	X_2, X_3 → DCMM_1; X_1 → DCMM_2	29.7	51.7	55.9	57.3	34.2
4	X_1, X_2, X_3 → separate DCMMs	29.3	51.1	55.2	57.0	33.7

4.4.3. Effectiveness of SCDFM

The SCDFM is added on top of the SIMM and DCMM to evaluate its impact on detection performance. As shown in row 4 of Table 2, compared with row 3, the model

achieves additional improvements of 2.1%, 2.4%, 1.8%, 2.0%, 2.2%, and 4.3% on APs1, APs2, APs3, APs4, APs, and AP, respectively. Similarly, as shown in row 4 of Table 3, compared with row 3, the model achieves additional improvements of 1.7%, 1.2%, 0.7%, 0.9%, 0.9%, and 0.4% on mAPs1, mAPs2, mAPs3, mAPs4, mAPs, and mAP, respectively. These results indicate that the SCDFM effectively integrates semantic and contextual information from preceding stages, leading to refined feature representations for small objects.

Notably, the SCDFM operates on the deepest layer from the backbone and the context-enhanced features generated by the DCMM. The SCDFM adaptively integrates the deep semantic and contextual information from the two sources, supplementing and balancing their information to facilitate comprehensive feature interactions.

4.4.4. Feature Map Visualization Analysis

To further verify the effectiveness of each module in enhancing feature representation, we visualize the feature activation maps generated by Grad-CAM, as illustrated in Figure 6.

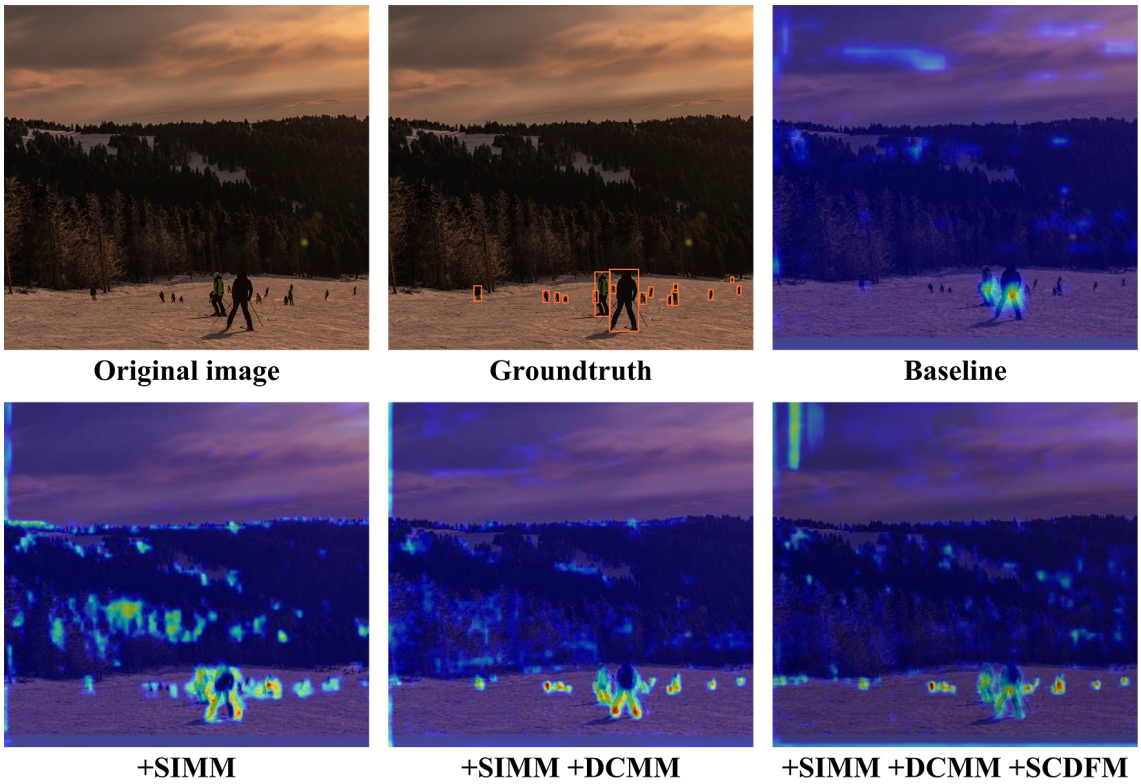

Figure 6. Feature map visualization of different module configurations.

The baseline exhibits sparse and scattered responses, focusing only on a few salient regions while missing many small-scale objects. After introducing the SIMM, the feature responses around the foreground are significantly enhanced, enabling the model to perceive the overall scene through semantic interaction modeling and expand the object search range. However, due to the absence of explicit local contextual constraints, residual activations can still be observed in some background regions. When the DCMM is integrated on top of the SIMM, the irrelevant responses in background areas are markedly suppressed, and the activations become more concentrated around the objects. This demonstrates that the DCMM effectively performs adaptive local context modeling, allowing the network to selectively strengthen foreground-relevant contextual cues while filtering redundant background information. With the inclusion of the SCDFM, the feature responses become more compact and discriminative, accurately focusing on the centers of small objects. This

confirms that the SCDFM refines the fused semantic and contextual features from the preceding stages, thereby enhancing the localization precision of small objects.

Overall, the progressive improvements brought by the three modules intuitively verify the synergistic effect of the SIMM, DCMM, and SCDFM, which together construct context-aware and semantically enriched feature representations for small object detection.

4.5. Comparison with State-of-the-Art Methods

To verify the performance of the CSIPN, we conduct comparative experiments against state-of-the-art methods on three UAV aerial datasets. All hyperparameters of the CSIPN are configured according to Section 4.2, while the other models adopt their original settings.

4.5.1. Quantitative Results

Table 5 compares the detection accuracy of our CSIPN with several mainstream methods on the VisDrone-DET validation set. In particular, the CSIPN shows remarkable advantages in small object detection, ranking first on mAPs1, mAPs2, mAPs3, mAPs4, and mAPs. Specifically, the CSIPN outperforms DINO (which ranks second on these metrics) by 1.9%, 2.8%, 2.5%, and 2.6% on mAPs1, mAPs2, mAPs3, and mAPs respectively, and exceeds YOLOv9-C (which ranks second on mAPs4) by 2.3% on mAPs4. From the perspective of the overall mAP, the CSIPN ranks second and leads the third-place DINO by 3.4%. These results demonstrate that the CSIPN not only has excellent adaptability in small object detection but also maintains high overall detection performance. Meanwhile, the CSIPN also achieves impressive parameter efficiency with a model size smaller than most of the methods listed in the table.

Table 5. Comparison results of detection accuracy between the proposed method and several state-of-the-art methods on the VisDrone validation set. Bold font indicates the best result of one metric.

Method	Backbone	mAPs1	mAPs2	mAPs3	mAPs4	mAPs	mAP	Params
RT-DETR [50]	ResNet-50	12.6	22.4	26.0	27.6	16.6	25.1	61.8M
DAB-DETR [51]	ResNet-50	10.7	22.3	23.2	32.0	17.0	28.0	43.7M
FCOS [11]	HRNetV2p-W32	17.7	28.8	30.9	41.2	24.7	34.7	37.1M
LDConv [52]	C3Net	23.4	29.7	24.1	29.8	25.9	33.9	**1.6M**
Deformable-DETR [53]	ResNet-50	21.5	32.8	36.6	42.6	28.1	36.9	40.8M
GFL [54]	ResNeXt-101	21.3	35.7	37.2	43.5	28.8	39.3	50.9M
ATSS [55]	ResNet-101	20.8	39.7	37.5	46.1	29.6	40.9	51.1M
DDOD [56]	ResNet-50	25.4	40.7	43.6	46.3	32.7	42.8	32.2M
GRoIE [57]	ResNet-50	25.9	43.4	48.2	52.8	34.5	45.1	43.3M
Double Heads [10]	ResNet-50	26.3	42.8	48.8	52.6	34.9	46.2	47.3M
PISA [58]	ResNeXt-101	26.7	44.2	47.7	55.9	35.7	47.0	60.0M
Cascade R-CNN [59]	HRNetV2p-W32	29.0	49.0	52.2	57.5	38.7	48.9	74.7M
Faster R-CNN [9]	HRNetV2p-W40	29.8	48.6	53.2	57.2	39.0	49.7	63.2M
VarifocalNet [60]	ResNet-101	31.6	46.9	49.5	54.4	39.4	48.9	51.7M
PConv [20]	C2fNet	34.1	47.9	50.5	54.5	40.7	48.7	3.0M
YOLOv10-M [61]	C2fNet	32.4	48.9	54.0	59.9	41.5	51.4	16.5M
EFC [62]	ResNet-18	35.0	49.7	50.8	58.8	42.6	51.1	40.0M
FFCA-YOLO [63]	C3Net	35.8	50.8	49.7	57.6	42.8	50.7	7.2M
YOLOv9-C [64]	GELAN	34.7	51.7	56.4	61.3	43.6	53.1	51.0M
Oriented-RCNN [65]	LEGNet-S [66]	37.0	52.8	56.7	59.9	45.2	**59.9**	29.8M
CentripetalNet [67]	HourglassNet-104	41.0	53.7	53.8	59.2	47.3	53.2	206.0M
DINO [48]	Swin-L	41.5	56.4	57.0	60.5	48.2	57.1	218.6M
CSIPN	C3Net	**43.4**	**59.2**	**59.5**	**63.6**	**50.8**	58.5	25.3M

To further demonstrate the superior performance of our method, we also conducted comparative experiments on our self-built WildDrone test set, and the results are shown

in Table 6. This dataset contains more diverse complex backgrounds and a larger number of tiny objects, making the detection task more challenging. The CSIPN achieves the best performance across all evaluation metrics and consistently outperforms other state-of-the-art methods. In terms of small object detection metrics, the CSIPN surpasses the second-best method PConv by 1.8%, 1.2%, 2.1%, 1.5%, and 2.0% on mAPs1, mAPs2, mAPs3, mAPs4, and mAPs, respectively. Regarding the overall detection performance, the CSIPN ranks second with an mAP of 68.4%. These comparative results once again indicate that the CSIPN has a significant advantage in handling detection tasks involving small-sized objects.

Table 6. Comparison results of detection accuracy between the proposed method and several state-of-the-art methods on the WildDrone test set. Bold font indicates the best result of one metric.

Method	Backbone	mAPs1	mAPs2	mAPs3	mAPs4	mAPs	mAP	Params
DAB-DETR [51]	ResNet-50	6.3	14.0	23.6	34.9	16.3	33.2	43.7M
FCOS [11]	HRNetV2p-W32	10.4	18.1	31.5	45.0	23.6	41.2	37.1M
LDConv [52]	C3Net	13.7	18.6	24.5	32.5	24.8	40.2	**1.6M**
Deformable-DETR [53]	ResNet-50	12.6	20.6	37.3	46.5	26.9	43.8	40.8M
GFL [54]	ResNeXt-101	12.5	22.4	37.9	47.5	27.5	46.6	50.9M
ATSS [55]	ResNet-101	12.2	24.9	38.2	50.3	28.3	48.5	51.1M
DDOD [56]	ResNet-50	14.9	25.5	44.4	50.6	31.3	50.8	32.2M
GRoIE [57]	ResNet-50	15.2	27.2	49.1	57.7	33.1	53.5	43.3M
Double Heads [10]	ResNet-50	15.4	26.8	49.7	57.4	33.4	54.8	47.3M
PISA [58]	ResNeXt-101	15.6	27.7	48.6	61.7	34.1	55.8	60.0M
RT-DETR [50]	ResNet-101	15.1	26.2	48.1	61.0	35.4	51.3	61.8M
Cascade R-CNN [59]	HRNetV2p-W32	17.0	30.7	53.1	62.8	37.0	58.4	74.7M
Faster R-CNN [9]	HRNetV2p-W40	17.4	30.5	54.2	62.5	37.3	59.3	63.2M
VarifocalNet [60]	ResNet-101	18.5	29.4	50.4	59.4	37.7	58.0	51.7M
YOLOv10-M [61]	C2fNet	19.3	30.7	55.0	65.4	39.7	61.7	16.5M
EFC [62]	ResNet-18	20.5	31.2	51.7	64.2	40.7	60.7	40.0M
YOLOv9-C [64]	GELAN	20.3	32.4	57.4	66.9	41.7	63.9	51.0M
Oriented-RCNN [65]	LEGNet-S [66]	21.6	33.1	57.7	65.4	43.2	**68.5**	29.8M
FFCA-YOLO [63]	C3Net	23.8	32.1	58.9	66.7	44.6	66.8	7.2M
CentripetalNet [67]	HourglassNet-104	24.0	33.7	54.8	64.6	45.2	63.1	206.0M
DINO [48]	Swin-L	24.3	35.4	58.6	66.1	46.1	67.8	218.6M
PConv [20]	C2fNet	24.4	36.6	60.5	69.4	46.3	67.7	3.0M
CSIPN	C3Net	**26.2**	**37.8**	**62.6**	**70.9**	**48.3**	68.4	25.3M

In addition, we perform comparative analyses with several representative detectors whose core mechanism is based on the self-attention architecture, including RT-DETR, DAB-DETR, Deformable-DETR, and DINO. As shown in Tables 5 and 6, our method achieves significantly higher accuracy in small object detection compared to these self-attention-based approaches. This demonstrates that beyond the global context modeling capability endowed by the SIMM, our method also benefits from the local context dynamic modeling capability introduced by the DCMM. These two complementary components enable our network to effectively capture both global dependencies and fine-grained local contextual cues, which is crucial for accurately detecting small and densely distributed objects. This advantage also explains why our model surpasses traditional self-attention-based detectors that primarily focus on global context modeling but are less effective in local context representation.

To further verify the feasibility of the CSIPN in practical applications, we evaluated its inference efficiency on a server equipped with an NVIDIA RTX 4090 GPU. Table 7 compares the small-object detection accuracy, number of parameters, giga floating-point operations (GFLOPs), and frames per second (FPS) of the CSIPN with several representative detectors. As shown, the CSIPN ranks second among all methods in computational

complexity (167 GFLOPs) and parameter count (25.3M), demonstrating its carefully designed lightweight architecture. The inference speed of the CSIPN reaches 91 FPS, which is significantly faster than Transformer-based and two-stage detectors, and it is only slightly lower than YOLOv10-M. These results indicate that our method achieves an ideal balance between accuracy and efficiency: it attains the highest detection accuracy (50.8% mAPs) while maintaining a lightweight design and fast inference speed. Although the proposed method has not yet been deployed on UAV hardware with limited computational resources, the experimental results sufficiently demonstrate its potential for real-time deployment.

Table 7. Comparison of inference efficiency of CSIPN and several representative methods on the VisDrone validation set. Bold font indicates the best result for each metric.

Method	Backbone	mAPs	Params	GFLOPs	FPS
RT-DETR	ResNet-50	16.6	61.8M	186	44
Deformable-DETR	ResNet-50	28.1	40.8M	**154**	23
GRoIE	ResNet-50	34.5	43.3M	494	19
Cascade R-CNN	HRNetV2p-W32	38.7	74.7M	254	22
Faster R-CNN	HRNetV2p-W40	39.0	63.2M	288	20
YOLOv10-M	C2fNet	41.5	**16.5M**	170	**105**
CentripetalNet	HourglassNet-104	47.3	206.0M	2500	8
DINO	Swin-L	48.2	218.6M	-	12
CSIPN	C3Net	**50.8**	25.3M	167	91

4.5.2. Qualitative Results

Figure 7 and Figure 8 present the qualitative results on the VisDrone-DET validation set and the WildDrone test set, respectively.

Several representative detection scenarios from the VisDrone-DET validation set are selected for visual comparison, as shown in Figure 7. As illustrated in regions #1 and #2, the baseline misses many small objects hidden in complex backgrounds, whereas the CSIPN successfully detects them despite strong background interference. This mainly benefits from the SIMM, which exploits interactive modeling between global semantic embeddings and shallow spatial descriptors to explicitly capture the latent relationships between small objects and backgrounds, thereby enhancing the network's ability to distinguish small objects from cluttered environments. In regions #3 and #4, the CSIPN performs better than the baseline in handling densely distributed small objects, thanks to the DCMM, which adaptively models context relevant to small objects so that the network can dynamically supplement the required contextual cues for different objects and improve detection capability in crowded scenes. These results demonstrate that our model exhibits strong scene adaptability in UAV aerial scenarios with complex backgrounds and dense small object distributions.

We further conduct qualitative comparison experiments on the WildDrone test set with several representative methods, including the powerful one-stage small object detector FFCA-YOLO and the recently proposed PConv designed for small objects, as shown in Figure 8. In region #1, both FFCA-YOLO and PConv miss small objects hidden in complex backgrounds, whereas the CSIPN effectively detects such objects and shows a clear advantage over other SOTA detectors. In region #2, even under poor illumination conditions, the CSIPN still demonstrates superior performance compared with other methods, mainly because the SCDFM dynamically fuses deep semantic features and shallow contextual details, thereby enhancing the network's scene adaptability. In region #3, FFCA-YOLO and PConv perform poorly in detecting densely distributed small objects, as they only expand the receptive fields to enhance feature representation but suffer from the limitation of fixed receptive fields, which makes them unsuitable for crowded scenes. Taken

together across regions #1, #2, and #3, the CSIPN adapts well to complex backgrounds and densely distributed objects, achieving excellent performance when detecting densely distributed person and effectively alleviating the severe missed detection problem of small objects. These results fully demonstrate that the CSIPN possesses both scene-adaptive and context-adaptive modeling capabilities, enabling strong accuracy for small object detection.

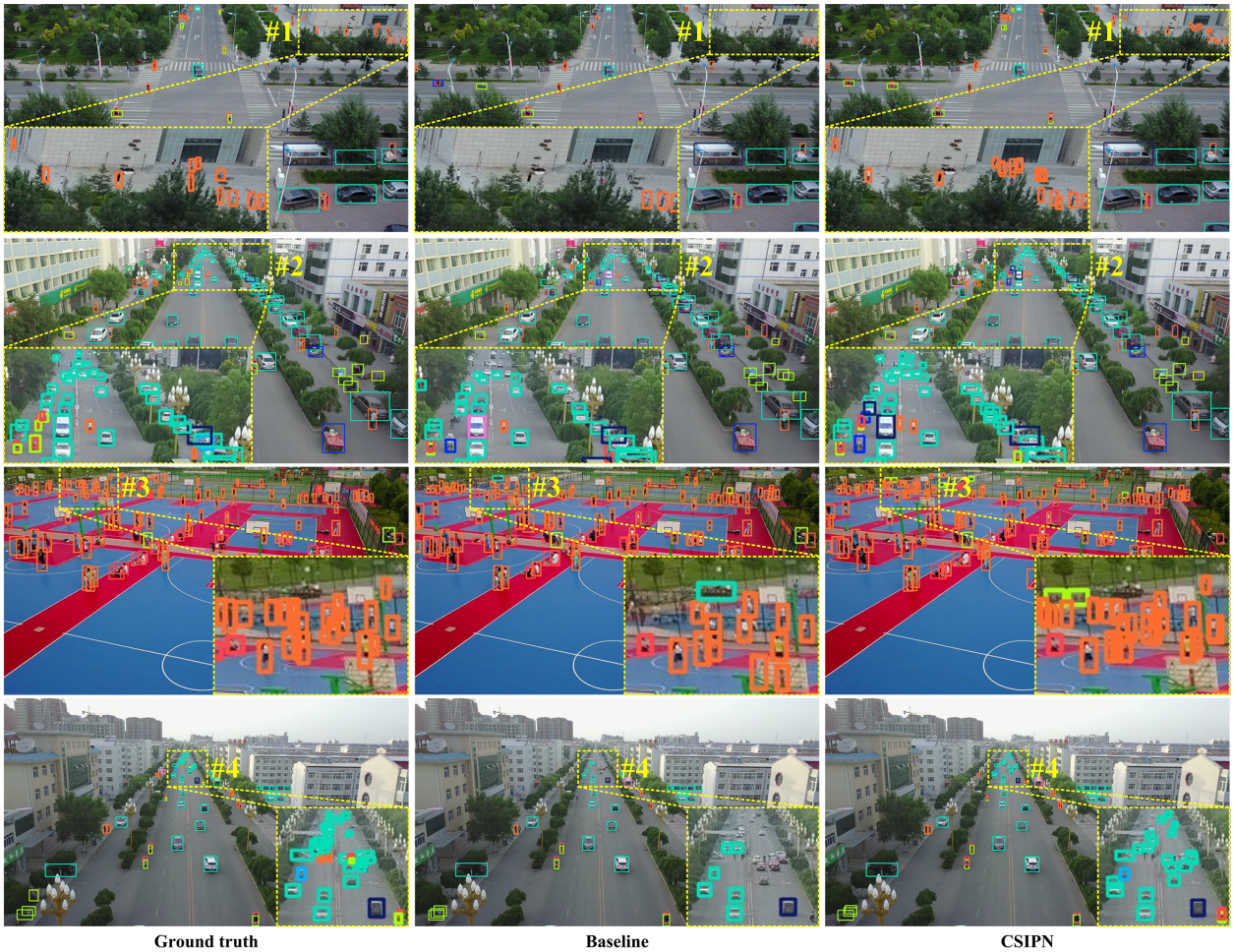

Figure 7. Qualitative comparison of CSIPN and the baseline on the VisDrone-DET validation set.

To further verify the robustness and scene adaptability of the proposed method, we conducted additional experiments under three challenging UAV task scenarios: foggy, nighttime, and strong-light conditions. As shown in Figure 9, our method consistently maintains superior detection performance compared with the baseline. In the foggy scene (region #1), the proposed model effectively detects small objects obscured by fog, whereas the baseline misses these objects. In the nighttime and strong-light scenes (regions #2 and #3), our method achieves significantly better performance than the baseline, accurately detecting small objects under varying illumination conditions. These results demonstrate that the proposed CSIPN can adapt to multiple extreme scenarios and exhibits strong robustness across diverse environmental conditions.

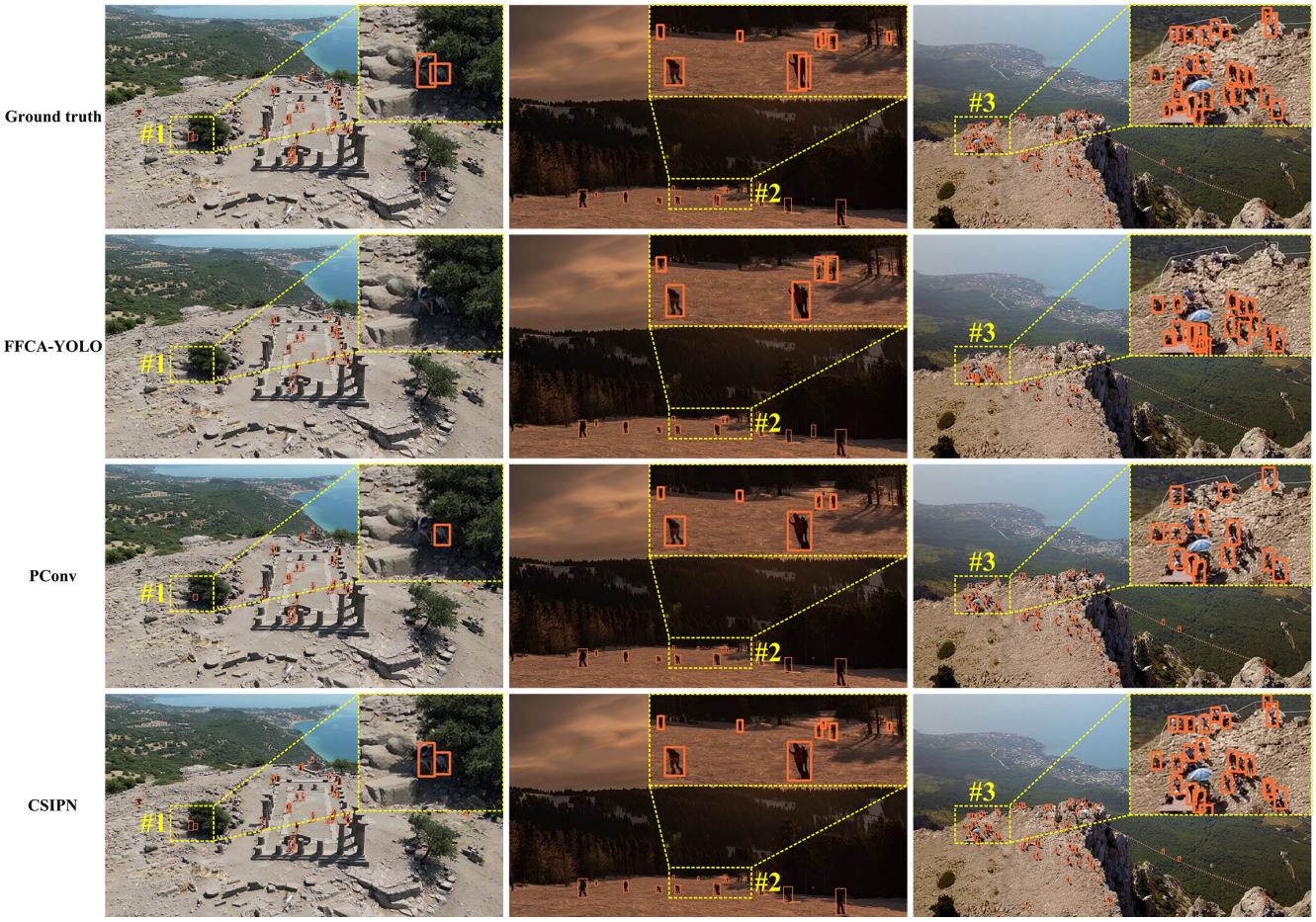

Figure 8. Qualitative comparison of CSIPN and some state-of-the-art methods on the WildDrone test set.

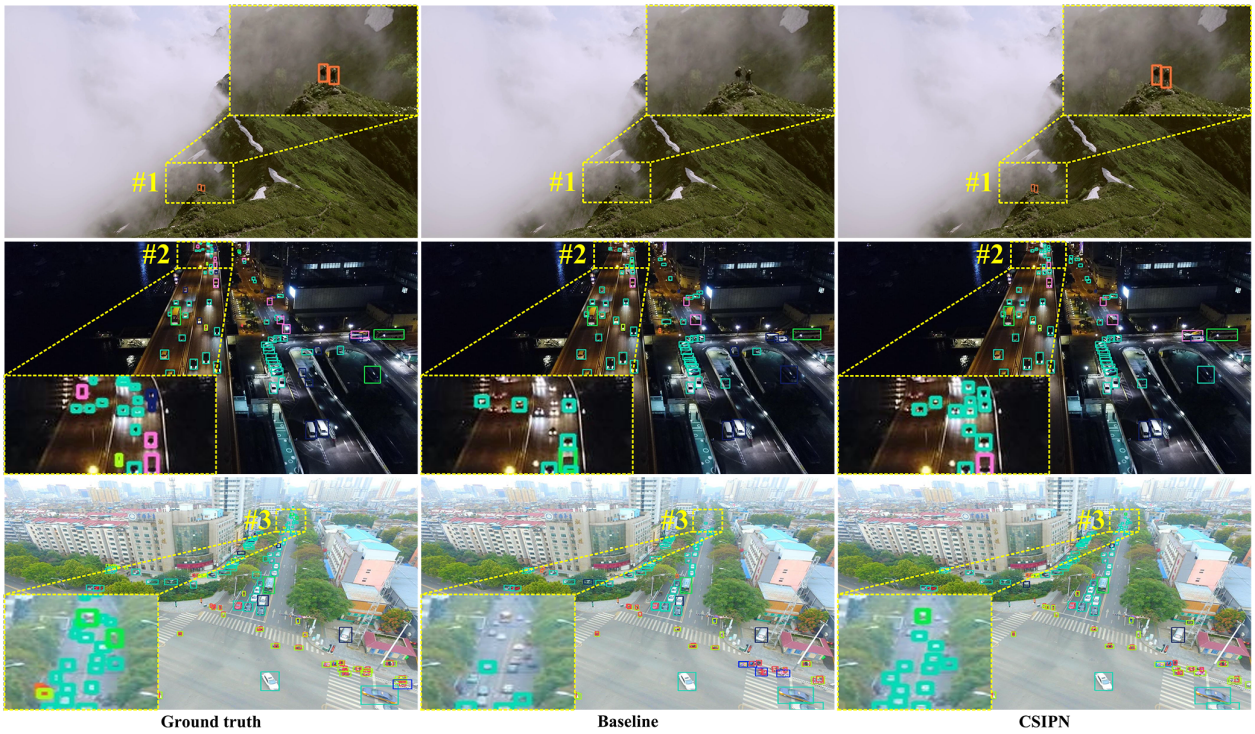

Figure 9. Visualization of robustness verification under extreme UAV conditions.

5. Discussion

Small object detection in UAV aerial images faces multiple challenges, including complex backgrounds, weak appearance features, densely distributed objects, and unstable illumination conditions. The proposed CSIPN achieves outstanding performance across multiple datasets through scene interaction modeling, dynamic context modeling, and semantic-context dynamic fusion. This section discusses and analyzes the advantages and underlying mechanisms of the CSIPN under different challenging scenarios based on experimental results.

5.1. Complex Background Interference and Weak Appearance Features

In UAV aerial scenes, small objects are often submerged in complex backgrounds, with extremely limited appearance cues, making them easily overlooked or misclassified by detectors. To address this problem, the proposed SIMM employs a lightweight self-attention mechanism to generate global scene semantic embedding vectors, which are interactively modeled with shallow spatial descriptors. This explicitly characterizes the latent relationships between small objects and the scene, selectively enhancing information relevant to small object detection while suppressing redundant background responses. Visualization results on the VisDrone-DET validation set (Figure 7) demonstrate that even in the presence of strong background interference such as buildings and vegetation, the CSIPN can accurately detect extremely small pedestrians and vehicles within the complex background.

5.2. Densely Distributed Small Objects and Insufficient Contextual Information

In certain UAV aerial scenes, small objects are often densely distributed at a large scale, with each object occupying very few pixels, making it difficult for the network to gather sufficient discriminative cues from a single receptive field. To address this, the proposed DCMM utilizes two dynamically adjustable receptive field branches to adaptively model contextual information relevant to small objects and dynamically supplement the specific contextual cues required by different objects. This design not only improves detection performance in densely distributed object scenes but also enhances the network's adaptability to different spatial layouts. Experimental results show that in densely distributed regions of the VisDrone-DET and WildDrone datasets (Figures 7 and 8), the CSIPN achieves a significantly lower missed detection rate compared to methods such as FFCA-YOLO and PConv, performing particularly well in detecting densely packed crowds at long distances.

5.3. Unstable Illumination and Scene Diversity

UAV aerial images are often affected by low illumination (e.g., captured at dusk or night), weather changes, and slight camera shake, all of which further weaken object features. During the semantic-context fusion stage, the SCDFM employs a dual weighting strategy in both channel and spatial dimensions to dynamically fuse deep semantic information with shallow contextual details. This not only preserves discriminative details but also enhances the global consistency of objects. The results on several challenging scenarios (Figure 9) show that the CSIPN surpasses other methods in both localization accuracy and stability, demonstrating strong robustness to interference and adaptability to diverse scenes.

6. Conclusions

To address the challenges of small object detection in UAV aerial scenarios, we propose a novel CSIPN, which significantly improves detection performance through scene interaction modeling, dynamic context modeling, and dynamic feature fusion. In the scene interaction modeling stage, the SIMM employs a lightweight self-attention mechanism

to generate a global semantic embedding and interact with shallow spatial descriptors, thereby enabling object–scene semantic interaction and achieving cross-scale alignment during feature fusion. In the dynamic context modeling stage, the DCMM adaptively models contextual information through two dynamically weighted receptive field branches, effectively supplementing the contextual cues required for detecting different small objects. In the semantic-context fusion stage, the SCDFM uses a dual channel-spatial weighting strategy to adaptively fuse deep semantic information with shallow contextual details, further optimizing the feature representation of small objects. Experiments on the TinyPerson dataset, the WAID dataset, the VisDrone-DET dataset, and our self-built WildDrone dataset show that the CSIPN achieves mAPs of 37.2%, 93.4%, 50.8%, and 48.3%, respectively, significantly outperforming existing state-of-the-art methods with only 25.3M parameters. Moreover, the CSIPN exhibits excellent robustness in challenging scenarios such as complex background occlusions and dense small object distributions, demonstrating its superiority in real-world applications. In the future, we will explore more efficient context modeling and lightweight feature fusion strategies to better support real-time deployment on resource-constrained UAV platforms.

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

Funding: This research was funded by the National Natural Science Foundation of China, grant number 62276204.

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

Acknowledgments: We thank the editors and reviewers for their hard work and valuable advice.

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

References

1. Doughty, C.L.; Cavanaugh, K.C. Mapping coastal wetland biomass from high resolution unmanned aerial vehicle (UAV) Imagery. *Remote Sens.* **2019**, *11*, 540. [CrossRef]
2. Dissanayaka, D.; Wanasinghe, T.R.; De Silva, O.; Jayasiri, A.; Mann, G.K.I. Review of navigation methods for UAV-based parcel delivery. *IEEE Trans. Autom. Sci. Eng.* **2024**, *21*, 1068–1082. [CrossRef]
3. Feroz, S.; Abu Dabous, S. UAV-based remote sensing applications for bridge condition assessment. *Remote Sens.* **2021**, *13*, 1809. [CrossRef]
4. Liu, J.; Liao, X.; Ye, H.; Yue, H.; Wang, Y.; Tan, X.; Wang, D. UAV swarm scheduling method for remote sensing observations during emergency scenarios. *Remote Sens.* **2022**, *14*, 1406. [CrossRef]
5. Liu, L.; Wang, A.; Sun, G.; Li, J.; Pan, H.; Quek, T.Q.S. Multi-objective optimization for data collection in UAV-assisted agricultural IoT. *IEEE Trans. Veh. Technol.* **2025**, *74*, 6488–6503. [CrossRef]
6. Jiang, L.; Yuan, B.; Du, J.; Chen, B.; Xie, H.; Tian, J.; Yuan, Z. MFFSODNet: Multiscale feature fusion small object detection network for UAV aerial images. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 1–14. [CrossRef]
7. Li, Y.; Li, Q.; Pan, J.; Zhou, Y.; Zhu, H.; Wei, H.; Liu, C. SOD-YOLO: Small-object-detection algorithm based on improved YOLOv8 for UAV images. *Remote Sens.* **2024**, *16*, 3057. [CrossRef]
8. Gui, S.; Song, S.; Qin, R.; Tang, Y. Remote sensing object detection in the deep learning era—A review. *Remote Sens.* **2024**, *16*, 327. [CrossRef]
9. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
10. Wu, Y.; Chen, Y.; Yuan, L.; Liu, Z.; Wang, L.; Li, H.; Fu, Y. Rethinking classification and localization for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10186–10195.

11. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.
12. Qu, J.; Tang, Z.; Zhang, L.; Zhang, Y.; Zhang, Z. Remote sensing small object detection network based on attention mechanism and multi-scale feature fusion. *Remote Sens.* **2023**, *15*, 2728. [CrossRef]
13. Zhu, Z.; Zheng, R.; Qi, G.; Li, S.; Li, Y.; Gao, X. Small object detection method based on global multi-level perception and dynamic region aggregation. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 10011–10022. [CrossRef]
14. Jiao, L.; Wang, M.; Liu, X.; Li, L.; Liu, F.; Feng, Z.; Yang, S.; Hou, B. Multiscale deep learning for detection and recognition: A comprehensive survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 5900–5920. [CrossRef]
15. Qi, G.; Zhang, Y.; Wang, K.; Mazur, N.; Liu, Y.; Malaviya, D. Small object detection method based on adaptive spatial parallel convolution and fast multi-scale fusion. *Remote Sens.* **2022**, *14*, 420. [CrossRef]
16. Zheng, X.; Qiu, Y.; Zhang, G.; Lei, T.; Jiang, P. ESL-YOLO: Small object detection with effective feature enhancement and spatial-context-guided fusion network for remote sensing. *Remote Sens.* **2024**, *16*, 4374. [CrossRef]
17. Wang, J.; Ma, M.; Huang, P.; Mei, S.; Zhang, L.; Wang, H. Remote sensing small object detection based on multicontextual information aggregation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 8248–8260. [CrossRef]
18. Chen, Z.; Ji, H.; Zhang, Y.; Liu, W.; Zhu, Z. Hybrid receptive field network for small object detection on drone view. *Chin. J. Aeronaut.* **2025**, *38*, 103127. [CrossRef]
19. Pan, X.; Ge, C.; Lu, R.; Song, S.; Chen, G.; Huang, Z.; Huang, G. On the integration of self-attention and convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 815–825.
20. Yang, J.; Liu, S.; Wu, J.; Su, X.; Hai, N.; Huang, X. Pinwheel-Shaped Convolution and Scale-Based Dynamic Loss for Infrared Small Target Detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; pp. 9202–9210. [CrossRef]
21. Du, D.; Zhu, P.; Wen, L.; Bian, X.; Lin, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y.; et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 213–226.
22. Yu, X.; Gong, Y.; Jiang, N.; Ye, Q.; Han, Z. Scale match for tiny person detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 1257–1265.
23. Xi, Y.; Jia, W.; Miao, Q.; Liu, X.; Fan, X.; Li, H. WAID: A Large-Scale Dataset for Wildlife Detection with Drones. *Appl. Sci.* **2023**, *13*, 10397. [CrossRef]
24. Deng, C.; Wang, M.; Liu, L.; Liu, Y.; Jiang, Y. Extended Feature Pyramid Network for Small Object Detection. *IEEE Trans. Multimed.* **2022**, *24*, 1968–1979. [CrossRef]
25. Xi, Y.; Jia, W.; Miao, Q.; Liu, X.; Fan, X.; Li, H. FiFoNet: Fine-Grained Target Focusing Network for Object Detection in UAV Images. *Remote Sens.* **2022**, *14*, 3919. [CrossRef]
26. Chen, Z.; Ji, H.; Zhang, Y.; Zhu, Z.; Li, Y. High-resolution feature pyramid network for small object detection on drone view. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 475–489. [CrossRef]
27. Chen, Y.; Ye, Z.; Sun, H.; Gong, T.; Xiong, S.; Lu, X. Global–Local Fusion With Semantic Information Guidance for Accurate Small Object Detection in UAV Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–15. [CrossRef]
28. Qu, J.; Liu, T.; Tang, Z.; Duan, Y.; Yao, H.; Hu, J. Remote sensing small object detection network based on multi-scale feature extraction and information fusion. *Remote Sens.* **2025**, *17*, 913. [CrossRef]
29. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
30. Min, K.; Lee, G.-H.; Lee, S.-W. Attentional feature pyramid network for small object detection. *Neural Netw.* **2022**, *155*, 439–450. [CrossRef]
31. Ge, Q.; Da, W.; Wang, M. MARFPNet: Multiattention and adaptive reparameterized feature pyramid network for small target detection on water surfaces. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 1–17. [CrossRef]
32. Kiobya, T.; Zhou, J.; Maiseli, B. A multi-scale semantically enriched feature pyramid network with enhanced focal loss for small-object detection. *Knowl.-Based Syst.* **2025**, *310*, 113003. [CrossRef]
33. Hu, J.; Li, Y.; Zhi, X.; Shi, T.; Zhang, W. Complementarity-Aware Feature Fusion for Aircraft Detection via Unpaired Opt2SAR Image Translation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–19. [CrossRef]
34. Liang, X.; Zhang, J.; Zhuo, L.; Li, Y.; Tian, Q. Small object detection in unmanned aerial vehicle images using feature fusion and scaling-based single shot detector with spatial context analysis. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 1758–1770. [CrossRef]
35. Cui, L.; Lv, P.; Jiang, X.; Gao, Z.; Zhou, B.; Zhang, L.; Shao, L.; Xu, M. Context-aware block net for small object detection. *IEEE Trans. Cybern.* **2022**, *52*, 2300–2313. [CrossRef]

36. Shi, T.; Gong, J.; Hu, J.; Zhi, X.; Zhu, G.; Yuan, B.; Sun, Y.; Zhang, W. Adaptive feature fusion with attention-guided small target detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–16. [CrossRef]
37. Hu, J.; Wei, Y.; Chen, W.; Zhi, X.; Zhang, W. CM-YOLO: Typical Object Detection Method in Remote Sensing Cloud and Mist Scene Images. *Remote Sens.* **2025**, *17*, 125. [CrossRef]
38. Hu, J.; Pang, T.; Peng, B.; Shi, Y.; Li, T. A Small Object Detection Model for Drone Images Based on Multi-Attention Fusion Network. *Image Vis. Comput.* **2025**, *155*, 105436. [CrossRef]
39. Li, Z.; Wang, Y.; Feng, H.; Chen, C.; Xu, D.; Zhao, T.; Gao, Y.; Zhao, Z. Local to Global: A Sparse Transformer-Based Small Object Detector for Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–16. [CrossRef]
40. Hu, P.; Ramanan, D. Finding tiny faces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 951–959.
41. Bai, Y.; Zhang, Y.; Ding, M.; Ghanem, B. SOD-MTGAN: Small object detection via multi-task generative adversarial network. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 206–221.
42. Rabbi, J.; Ray, N.; Schubert, M.; Chowdhury, S.; Chao, D. Small-object detection in remote sensing images with end-to-end edge-enhanced GAN and object detector network. *Remote Sens.* **2020**, *12*, 1432. [CrossRef]
43. Bashir, S.M.A.; Wang, Y. Small object detection in remote sensing images with residual feature aggregation-based super-resolution and object detector network. *Remote Sens.* **2021**, *13*, 1854. [CrossRef]
44. Zhang, J.; Lei, J.; Xie, W.; Fang, Z.; Li, Y.; Du, Q. SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–15. [CrossRef]
45. Liu, J.; Zhang, J.; Ni, Y.; Chi, W.; Qi, Z. Small-object detection in remote sensing images with super-resolution perception. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 15721–15734. [CrossRef]
46. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The PASCAL visual object classes (VOC) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
47. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
48. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.-Y. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv* **2022**, arXiv:2203.03605.
49. Jocher, G. Ultralytics YOLO11. Available online: https://github.com/ultralytics/ultralytics (accessed on 2 March 2025).
50. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs beat YOLOs on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2024; pp. 16965–16974.
51. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. Dab-DETR: Dynamic anchor boxes are better queries for DETR. *arXiv* **2022**, arXiv:2201.12329.
52. Zhang, X.; Song, Y.; Song, T.; Yang, D.; Ye, Y.; Zhou, J.; Zhang, L. LDConv: Linear deformable convolution for improving convolutional neural networks. *Image Vis. Comput.* **2024**, *149*, 105190. [CrossRef]
53. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
54. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21002–21012.
55. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9759–9768.
56. Chen, Z.; Yang, C.; Li, Q.; Zhao, F.; Zha, Z.-J.; Wu, F. Disentangle your dense object detector. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, China, 20–24 October 2021; pp. 4939–4948.
57. Rossi, L.; Karimi, A.; Prati, A. A novel region of interest extraction layer for instance segmentation. In Proceedings of the 2020 25th International Conference on Pattern Recognition, Milan, Italy, 10–15 January 2021; pp. 2203–2209.
58. Cao, Y.; Chen, K.; Loy, C.C.; Lin, D. Prime sample attention in object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11583–11591.
59. Cai, Z.; Vasconcelos, N. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 1483–1498. [CrossRef] [PubMed]
60. Zhang, H.; Wang, Y.; Dayoub, F.; Sunderhauf, N. Varifocalnet: An IoU-aware dense object detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8514–8523.
61. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. Yolov10: Real-time end-to-end object detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 107984–108011.
62. Xiao, Y.; Xu, T.; Yu, X.; Fang, Y.; Li, J. A lightweight fusion strategy with enhanced interlayer feature correlation for small object detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–11. [CrossRef]

63. Zhang, Y.; Ye, M.; Zhu, G.; Liu, Y.; Guo, P.; Yan, J. FFCA-YOLO for small object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–15. [CrossRef]
64. Wang, C.-Y.; Yeh, I.-H.; Mark Liao, H.-Y. Yolov9: Learning what you want to learn using programmable gradient information. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 27–31 October 2024; pp. 1–21.
65. Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented R-CNN for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 3520–3529.
66. Lu, W.; Chen, S.B.; Li, H.D.; Shu, Q.-L.; Ding, C.H.Q.; Tang, J.; Luo, B. LegNet: Lightweight edge-Gaussian driven network for low-quality remote sensing image object detection. *arXiv* **2025**, arXiv:2503.14012.
67. Dong, Z.; Li, G.; Liao, Y.; Wang, F.; Ren, P.; Qian, C. Centripetalnet: Pursuing high-quality keypoint pairs for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 10519–10528.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.