

FRM-hm: Heatmap Refinement for Small Object Detection in Center-based Detectors with Application to Rice-field Bird Monitoring

Kassi Benoît Yao ^{1,*}, Armand Kodjo Atiampo ², and Georges Laussane Loum ¹

¹ École Doctorale Polytechnique des Sciences et Techniques de l'Ingénieur (EDP-STI), UMRI Mathématiques et Sciences du Numérique, Institut National Polytechnique Félix Houphouët-Boigny, Yamoussoukro, Côte d'Ivoire

² Informatique, UFR Informatique et Sciences du Numérique, Université Virtuelle de Côte d'Ivoire, Abidjan, Côte d'Ivoire

Email: kassi.yao22@inphb.ci (K.B.Y.); armand.atiampo@uvci.edu.ci (A.K.A.); georges.loum@inphb.ci (G.L.L.)

*Corresponding author

Abstract—The automatic detection of granivorous birds in images acquired in rice-field environments is an important challenge for crop protection, yet it remains difficult due to the small object size, scale variations, partial occlusions, and the complexity of natural scenes. We propose REBDetX01, an object detection model based on the CenterNet architecture, integrating a dedicated refinement module in the heatmap prediction branch, referred to as FRM-hm. This module combines a channel recalibration mechanism and dilated spatial attention to improve the organization of heatmap activations and the localization of object centers, without modifying the network backbone or exploiting temporal information. The model is evaluated on the FBD-SV-2024 dataset, characterized by predominantly small flying bird instances and complex backgrounds. Experimental results show that REBDetX01 outperforms the CenterNet baseline, improving Average Precision (AP) from 63.36% to 65.52% (+2.16 points), AP50 from 86.07% to 87.90% (+1.83 points), and Average Precision for small objects (APS) from 61.84% to 63.03% (+1.19 points), thereby indicating its effectiveness for small-object detection. Ablation studies further show that these gains arise from the complementarity between the channel recalibration and dilated spatial attention mechanisms integrated within FRM-hm. Overall, the results highlight the benefit of targeted heatmap refinement for detecting challenging objects in complex natural scenes.

Keywords—rice-field environments, object detection, small objects, CenterNet model, attention mechanism

I. INTRODUCTION

Protecting rice fields from granivorous birds is a major challenge for food security and the economic sustainability of rice production systems, particularly in regions where rice cultivation represents a key agricultural activity [1–4]. Large flocks of granivorous birds can cause significant yield losses, especially during the sensitive stages of rice heading and grain maturation [4, 5].

To mitigate these losses, various bird deterrence techniques have traditionally been employed, including manual scaring, acoustic devices, and visual deterrents. However, these approaches are often costly, labor-intensive, and only temporarily effective, as birds rapidly adapt to repetitive deterrence mechanisms [6–10].

Consequently, the development of automatic monitoring and detection systems has emerged as a promising alternative for improving crop protection strategies.

Recent advances in computer vision and deep learning have enabled the development of automated systems capable of analyzing images captured in natural environments to detect objects of interest [11, 12]. Despite these advances, detecting granivorous birds in rice-field imagery remains particularly challenging. Agricultural scenes typically exhibit high visual complexity, including heterogeneous backgrounds, dense vegetation, illumination variations, and low contrast between birds and their surroundings. In addition, birds often appear as very small objects in images, and their appearance may vary significantly due to scale changes, motion, and partial occlusions [13].

These characteristics considerably limit the performance of generic object detection models, particularly when accurate localization of small objects is required. Traditional anchor-based detectors often suffer from an imbalance between positive and negative training samples, as well as a strong sensitivity to hyperparameter choices such as anchor sizes and aspect ratios [14, 15]. These limitations have motivated the development of anchor-free detection methods, which aim to simplify the detection formulation and improve robustness to object scale variations. Among these approaches, methods based on keypoint detection have gained increasing attention for object localization tasks [16].

In this category, CenterNet formulates object detection as a problem of predicting object centers represented as

local maxima on a heatmap [16]. By focusing on center-point localization, CenterNet eliminates the need for predefined anchor boxes and offers improved robustness to scale variations. However, the performance of center-based detectors depends heavily on the quality of the predicted heatmap, which serves as the primary representation for object center localization [16]. In complex scenes such as rice-field environments, visual noise, repetitive textures, and the presence of small objects may degrade the quality of this heatmap, resulting in inaccurate localization or missed detections [13].

In this work, we hypothesize that targeted refinement of the heatmap prediction branch can improve the performance of center-based object detectors in the specific context of granivorous bird detection. Instead of introducing global attention mechanisms applied uniformly across the entire network, we propose a lightweight and localized attention module designed specifically for the heatmap prediction branch. This module, called the Feature Refinement Module for heatmap (FRM-hm), aims to enhance the detector's ability to highlight regions corresponding to object centers while preserving the local maxima necessary for accurate localization. This design follows the general principle of attention mechanisms in Convolutional Neural Networks (CNNs), while distinguishing itself by focusing on the most critical component of center-based detection architectures [17, 18].

The integration of the proposed FRM-hm module into the CenterNet architecture yields a detection model referred to as Rice Eating Bird Detector X01 (REBDetX01). This model is specifically designed to address the challenges of detecting granivorous birds in rice-field images while preserving the original detection formulation and avoiding significant additional computational cost.

The main methodological contribution of this work is the design of FRM-hm, a dedicated heatmap refinement module for center-based object detection. Unlike generic attention blocks applied uniformly across the network, FRM-hm is inserted only into the heatmap prediction branch, which is the component directly responsible for object-center localization. The complete model, referred to as REBDetX01, is obtained by integrating FRM-hm into a CenterNet-based detector while keeping the backbone, size branch, and offset branch unchanged.

The proposed approach has four main strengths. First, FRM-hm targets the heatmap branch, which is the most localization-sensitive component in a center-based detector. Second, the proposed module preserves the original backbone as well as the size and offset branches, making the modification localized and easy to analyze. Third, the experimental results show consistent improvements over the CenterNet baseline on AP, AP50, and APS, with particular relevance for small bird detection. Fourth, the ablation study and comparison with Squeeze-and-Excitation Block (SE-Block) [17] and Convolutional Block Attention Module (CBAM) [18] confirm that the observed gains are associated with the

specific design of FRM-hm rather than with the generic use of attention.

The main contributions of this paper are summarized as follows:

- We design FRM-hm, a lightweight attention module specifically dedicated to the heatmap branch of a CenterNet-based object detector;
- We integrate FRM-hm into a center-based detection architecture, resulting in the proposed model REBDetX01 for detecting granivorous birds in rice-field images.

The remainder of this paper is organized as follows. Section II reviews related work on bird detection, small object detection, and attention mechanisms in convolutional neural networks. Section III presents the proposed methodology, including the design of the FRM-hm module and its integration into CenterNet. Section IV describes the experimental framework and reports quantitative and qualitative results. Finally, Section V concludes the paper and discusses possible directions for future research.

II. LITERATURE REVIEW

A. Bird Detection in Natural Environments

Automatic bird detection in natural environments has attracted increasing attention in recent years due to its relevance for ecological monitoring, biodiversity studies, and agricultural crop protection [6, 7, 19]. In agricultural contexts, and particularly in rice-growing areas, the detection of granivorous birds presents several specific challenges related to the small size of the targets, their high mobility, and the complexity of the surrounding environment.

Natural agricultural scenes typically contain heterogeneous backgrounds composed of vegetation, soil textures, and sky regions, often combined with strong illumination variations. Under these conditions, birds may appear as small and weakly contrasted objects, which makes their detection difficult for generic object detectors originally designed for structured scenes or datasets dominated by medium and large objects [13, 20, 21].

To address this problem, Sun *et al.* [13] introduced the FBD-SV-2024 dataset, which is specifically designed for bird detection in natural surveillance environments. The dataset contains images and video frames captured in agricultural fields and includes numerous instances of birds appearing as extremely small objects. According to the authors, many birds in the dataset occupy only a few tens of pixels in the image and often appear in visually cluttered scenes. These characteristics make the dataset particularly suitable for evaluating detection models under realistic and challenging conditions.

Recent years have also seen the introduction of several bird detection datasets and methods designed for challenging acquisition conditions. Drone2021 addresses distant bird detection from a drone perspective [22], while the MVA2023 small object detection challenge focuses on spotting birds as small objects [23]. AirBirds was introduced for bird strike prevention in real-world airport

environments [24]. More recently, FBD-SV-2024 was proposed for flying bird detection in surveillance videos and includes objects with inconspicuous single-frame features, small sizes, and significant shape variability [13]. In addition, Flying Bird Object Detection based on Motion Information (FBOD-BMI) and Flying Bird Object Detection for Surveillance Video (FBOD-SV) were proposed as bird-specific video detection methods exploiting motion or spatio-temporal information to improve detection in surveillance videos [25, 26].

These observations confirm that the detection of granivorous birds in rice-field environments represents a representative and demanding case of small object detection in complex natural scenes.

B. Small Objects Detection

Small object detection is widely recognized as one of the most challenging problems in computer vision [20]. In deep convolutional detection architectures, successive convolution and downsampling operations progressively reduce the spatial resolution of feature maps. While this process increases the receptive field and improves high-level semantic representation, it may also lead to the partial or complete loss of discriminative information associated with small objects [16, 20].

This difficulty is exacerbated when small objects appear with low contrast or are partially occluded by elements of the background, such as vegetation or terrain structures in agricultural scenes [13]. As a result, the visual cues necessary for reliable object detection may become difficult to preserve throughout the feature extraction pipeline.

Standard evaluation protocols such as the Microsoft COCO benchmark explicitly distinguish small objects as a separate evaluation category [20]. Numerous studies have shown that detection performance on small objects remains significantly lower than that achieved on medium or large objects across most detection architectures. These limitations have motivated extensive research efforts aimed at improving multi-scale feature representation, feature fusion strategies, and detection head designs capable of preserving fine spatial details.

Recent object detection studies continue to show that small-object detection remains difficult because fine spatial cues are progressively weakened during feature extraction and downsampling. Modern detectors address this issue through different strategies, including multi-scale feature fusion, high-resolution feature preservation, decoupled prediction heads, transformer-based decoding, and task-specific detection heads. These developments highlight the need to analyze not only global detector architectures, but also the specific components responsible for localization.

C. Anchor-based Detectors and Their Limitations for Small Objects

Anchor-based detectors constitute one of the most widely used families of object detection models. Prominent examples include architectures such as Faster Region-based Convolutional Neural Network (Faster R-CNN), Single Shot MultiBox Detector (SSD), and You

Only Look Once (YOLO) [11, 12, 27]. These methods rely on predefined sets of anchor boxes with different scales and aspect ratios that serve as reference templates for bounding box regression.

Although anchor-based detectors have achieved strong performance in many applications, they present several limitations when applied to the detection of small objects [20, 27]. A major difficulty lies in designing appropriate anchor configurations. The selection of appropriate anchor sizes and aspect ratios is highly dependent on the characteristics of the dataset, and poorly configured anchors may lead to suboptimal detection performance [12, 27].

Another limitation arises from the low overlap between anchor boxes and small objects. Because small objects occupy only a limited number of pixels, the probability that predefined anchors sufficiently overlap with the ground truth bounding boxes is often low. This situation leads to a strong imbalance between positive and negative training samples, which complicates the learning process and may reduce training stability [12, 27].

These issues are particularly problematic in dense scenes such as rice-field environments, where numerous small objects may appear simultaneously in cluttered backgrounds [13].

D. Anchor-free Detection and Center-based Localization

To address these limitations, several studies have proposed anchor-free detection methods, which aim to simplify the detection formulation and reduce dependency on predefined anchor configurations.

Among these approaches, keypoint-based detection methods have gained increasing attention, particularly for handling scale variability in object detection tasks [16]. These methods detect specific points associated with objects rather than relying on predefined bounding box templates.

In this category, CenterNet reformulates object detection as a problem of locating object centers represented as local maxima on a heatmap [16]. The model predicts a heatmap indicating the probability of each spatial location being the center of an object. Additional regression heads are then used to estimate bounding box sizes and positional offsets.

This formulation offers several advantages. First, it eliminates the need for anchor boxes, thereby reducing the sensitivity of the detection pipeline to anchor configuration. Second, it explicitly separates the problem of object center localization from bounding box size estimation.

However, as highlighted by the authors of CenterNet, the performance of the model depends critically on the quality of the predicted heatmap, which serves as the primary representation for object center localization [16]. In complex scenes and for small objects, visual noise and background structures may degrade the clarity of heatmap activations, resulting in inaccurate localization or missed detections.

Recent one-stage and anchor-free detectors such as YOLOX [28], as well as recent YOLO-based models such as YOLOv9 [29] and YOLOv10 [30], have improved the

speed-accuracy trade-off through architectural refinements, improved assignment strategies, and more efficient prediction heads. In parallel, transformer-based detectors such as Real-Time Detection Transformer (RT-DETR) [31] have introduced real-time end-to-end detection pipelines. These models represent important advances in modern object detection. However, their contributions are mainly detector-level innovations, whereas the present work focuses on a heatmap-specific refinement strategy within a center-based detection framework.

E. Attention Mechanisms in Convolutional Neural Networks (CNN)

Attention mechanisms have been widely explored to improve the ability of CNNs to focus on the most informative features within feature maps. One of the most influential approaches is the Squeeze-and-Excitation (SE) block, which introduces channel-wise attention by dynamically recalibrating feature maps according to their global importance [17].

Another widely used mechanism is the Convolutional Block Attention Module (CBAM), which sequentially applies channel attention and spatial attention to refine feature representations [18]. These attention mechanisms have demonstrated effectiveness in various tasks, including image classification, object detection, and semantic segmentation.

Despite their success, most attention modules are typically applied globally or uniformly throughout the network architecture [17, 18]. In the context of small object detection, such global application may sometimes produce undesirable effects. In particular, excessive smoothing of feature maps may weaken the local maxima that are essential for accurate object center localization in keypoint-based detectors [16].

These observations suggest that attention mechanisms may be more effective when carefully positioned within specific components of the detection architecture, particularly those directly responsible for localization.

F. Discussion and Positioning of the Proposed Approach

The review of existing work highlights several major challenges associated with the detection of granivorous birds in rice-field environments. These challenges include the small size of the objects, the visual complexity of natural scenes, and the strong dependency of detection performance on accurate localization.

Anchor-free detectors such as CenterNet provide a promising framework for addressing these constraints due to their center-based detection formulation. However, their performance remains strongly dependent on the quality of the predicted heatmap.

These observations motivate the design of a lightweight and localized attention module specifically dedicated to the heatmap prediction branch. Such a module aims to improve the discriminative structure of heatmap activations while preserving the local maxima necessary for accurate object center localization.

In light of recent progress in object detection, the proposed work should be understood as complementary to

modern State-of-the-Art (SOTA) detectors. Rather than proposing a new general-purpose detector family, this study investigates whether targeted refinement of the heatmap branch can improve object-center localization for small birds in complex natural scenes. This positioning is particularly relevant for center-based detectors, where heatmap quality directly affects localization accuracy.

This work does not aim to introduce a new general-purpose detector competing directly with all recent detector families. Instead, it focuses on a specific limitation of center-based detection, namely the sensitivity of object-center localization to the quality of the predicted heatmap. CenterNet was therefore selected as a controlled anchor-free center-based framework in which the effect of a heatmap-specific refinement module can be isolated and analyzed. Recent detectors such as YOLOX, YOLOv9, YOLOv10, RT-DETR, and other end-to-end or Non-Maximum Suppression-free (NMS-free) architectures improve detection through different mechanisms, including decoupled prediction heads, improved assignment strategies, transformer-based decoding, or removal of post-processing bottlenecks. These approaches are important references for modern object detection. However, the contribution of this work is complementary: it investigates whether a localized refinement of the heatmap branch can improve small-object localization in a center-based detector.

The strength of the proposed approach lies in its targeted nature. Instead of modifying the entire detector or introducing a new general-purpose detection architecture, FRM-hm focuses on the heatmap representation used for object-center localization. This design is particularly relevant for small bird detection, where weak visual cues and complex backgrounds may degrade the quality of heatmap peaks.

The following section presents the proposed methodology, including the design of the FRM-hm module and its integration into the REBDetX01 detection model.

III. MATERIALS AND METHODS

A. CenterNet Formulation in a Single-class Discrete Detection Setting

1) Detection principle

REBDetX01 is built upon CenterNet, an anchor-free detector that models each object as the center point of its bounding box [16]. In the present study, the problem is formulated as a single-class discrete detection task, since only one object category is considered, namely granivorous birds. Under this setting, each bird instance is represented by one center point and one bounding box.

Let $I \in \mathbb{R}^{C \times H \times W}$ denote an input RGB image, where W and H are the image width and height, respectively. Following CenterNet [16], the detector aims to predict a keypoint heatmap

$$\hat{Y} \in [0, 1]^{C \times \frac{W}{r} \times \frac{H}{r}} \quad (1)$$

where r is the output stride and C is the number of object categories. In the present work, $C = 1$, since only one category is considered.

In addition to the heatmap, CenterNet predicts a size map and a local offset map:

$$\hat{S} \in \mathbb{R}^{2 \times \frac{W}{r} \times \frac{H}{r}}, \hat{O} \in \mathbb{R}^{2 \times \frac{W}{r} \times \frac{H}{r}} \quad (2)$$

Thus, the detector outputs, at each spatial location, one center-confidence score, one 2D size vector, and one 2D offset vector [16].

Fig. 1 shows the overall architecture of the proposed REBDetX01 model and highlights that FRM-hm is only inserted at the entrance of the heatmap head, while the size and offset heads remain identical to those of the baseline CenterNet architecture.

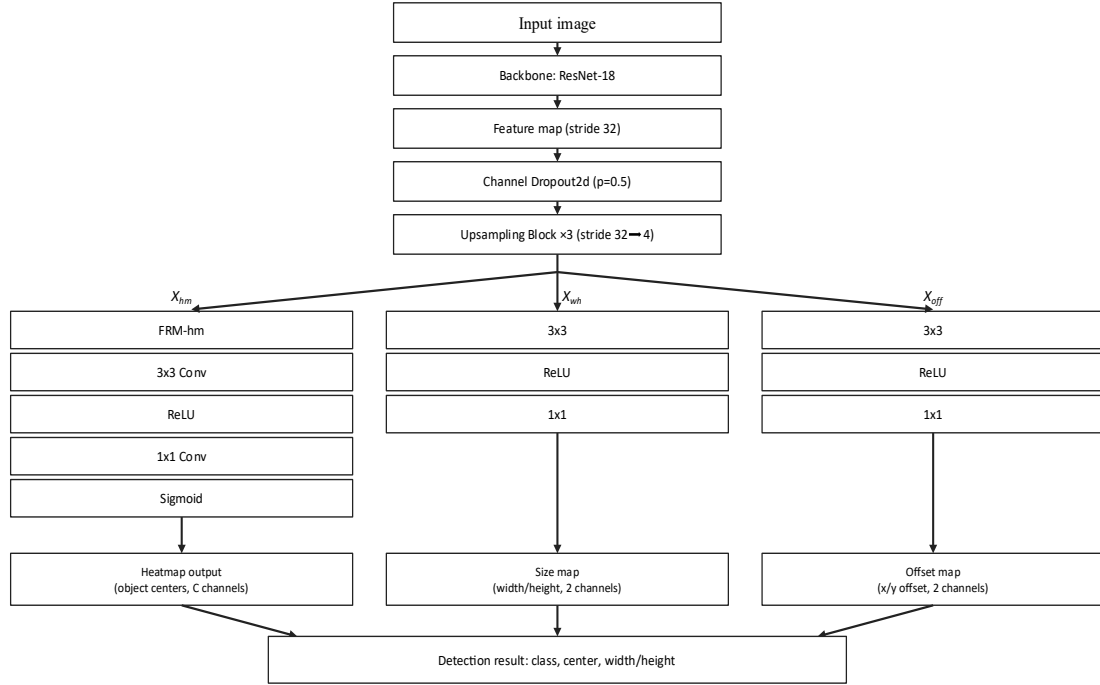


Fig. 1. General architecture of REBDetX01 and positioning of the FRM-hm module.

2) Heatmap-based object center representation

For each ground-truth object k , let b_k its bounding box be defined as

$$b_k = (x_1, y_1, x_2, y_2) \quad (3)$$

where (x_1, y_1) and (x_2, y_2) denote the top-left and bottom-right corners of the bounding box, respectively. Its center point is

$$p_k = \left(\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2} \right) \quad (4)$$

Following CenterNet [16], the low-resolution center is obtained by

$$\tilde{p}_k = \left\lfloor \frac{p_k}{r} \right\rfloor \quad (5)$$

The ground-truth heatmap Y is generated by placing a Gaussian kernel at each projected center:

$$Y_{xyc} = \exp \left(-\frac{(x - \tilde{p}_x)^2 + (y - \tilde{p}_y)^2}{2\sigma_p^2} \right) \quad (6)$$

where σ_p is an object-size-adaptive standard deviation [16]. If two Gaussians of the same class overlap, the element-wise maximum is retained [16]. In the present single-class setting, the heatmap has only one channel, and each spatial location estimates the probability that it corresponds to a bird center.

3) Offset and size regression

To recover the discretization error caused by the output stride, CenterNet predicts a local offset for each center point [16]. The target offset for object k is

$$o_k = \frac{p_k}{r} - \tilde{p}_k \quad (7)$$

In parallel, the detector regresses the object size

$$s_k = (x_2 - x_1, y_2 - y_1) \quad (8)$$

At inference time, object centers are first extracted as local peaks from the predicted heatmap. The corresponding offset vectors compensate for coordinate quantization, while the size vectors provide the final bounding-box dimensions [16]. This decomposition is particularly suitable for center-based detection, where localization accuracy depends directly on the quality of heatmap peaks.

4) Detection heads and loss functions

In CenterNet, all outputs share a common fully convolutional backbone. For each modality, the backbone features are then passed through a separate 3×3 convolution, a ReLU activation, and a final 1×1 convolution [16]. In the present work, this principle is preserved.

The heatmap is supervised with the penalty-reduced focal loss used in CenterNet [16]:

$$L_k = -\frac{1}{N} \sum_{x,y,c} \begin{cases} (1 - \hat{Y}_{xyc})^\alpha \log(\hat{Y}_{xyc}), & \text{if } Y_{xyc} = 1 \\ (1 - Y_{xyc})^\beta (\hat{Y}_{xyc})^\alpha \log(1 - \hat{Y}_{xyc}), & \text{otherwise} \end{cases} \quad (9)$$

where N is the number of objects in the image. As in CenterNet, $\alpha = 2$ and $\beta = 4$ are used [16].

The local offset is optimized with an L1 loss evaluated only at the center locations:

$$L_{off} = \frac{1}{N} \sum_{k=1}^N |\hat{p}_k - (\frac{p_k}{r} - \tilde{p}_k)| \quad (10)$$

Similarly, the size branch is trained with an L1 loss at the object centers:

$$L_{size} = \frac{1}{N} \sum_{k=1}^N |\hat{s}_k - s_k| \quad (11)$$

The overall detection objective is therefore

$$L_{det} = L_k + \lambda_{size} L_{size} + \lambda_{off} L_{off} \quad (12)$$

where $\lambda_{size} = 0.1$ and $\lambda_{off} = 1$, following the original CenterNet formulation [16].

This decomposition highlights the distinct role of each head. The heatmap head determines where object centers are located. The size and offset heads determine how these centers are converted into bounding boxes. Consequently, any degradation in heatmap quality directly affects localization performance. This observation motivates the refinement strategy proposed in Section III-B.

B. FRM-hm Module and its Integration into CenterNet

1) Motivation and insertion point

The central hypothesis of this work is that the overall performance of a center-based detector depends strongly on the discriminative quality of the heatmap representation. This issue is particularly critical on FBD-SV-2024 [13], where bird objects are often small and embedded in visually complex backgrounds. The dataset study reports that flying bird objects are mainly distributed between 10×10 and 70×70 pixels, accounting for about 94% of all objects, and that about 49.99% of them are smaller than 32×32 pixels.

To address this issue, the proposed Feature Refinement Module for heatmap (FRM-hm) is inserted only in the heatmap branch. More precisely, FRM-hm is placed at the entrance of the heatmap head, that is, before the

convolutional block used to compute the heatmap prediction, while the size and offset heads remain unchanged. This formulation is consistent with the architecture described in this work and with the implementation, where the shared upsampled feature map is first refined by FRM-hm and only then passed to the heatmap head.

FRM-hm is specifically designed for the heatmap branch of a center-based detector. It combines channel recalibration, dilated spatial attention, and residual refinement to improve the structuring of heatmap activations before final center prediction. The channel recalibration step emphasizes feature channels that are more informative for object-center localization. The dilated spatial attention step then enhances spatially coherent responses around potential object centers while incorporating limited contextual information. Finally, residual refinement preserves local corrective information that may be useful for generating sharp and stable heatmap peaks. This targeted design distinguishes FRM-hm from standard attention modules such as SE-Block or CBAM, which are generally applied as generic feature refinement blocks rather than as heatmap-specific localization modules.

Let F denote the shared upsampled feature map produced by the backbone and decoder. In the proposed model, the tensor feeding the heatmap branch is explicitly defined as

$$X_{hm} = F \quad (13)$$

2) Global formulation of FRM-hm

FRM-hm combines three successive operations: channel recalibration, dilated spatial attention, and residual refinement. Its output is written as

$$\tilde{X}_{hm} = FRMhm(X_{hm}) \quad (14)$$

The refined tensor $\tilde{X}_{hm}$ is then processed by the standard heatmap head:

$$Z_{hm} = \text{ReLU}(\text{Conv}_{3 \times 3}(\tilde{X}_{hm})) \quad (15)$$

$$\hat{Y} = \sigma(\text{Conv}_{1 \times 1}(Z_{hm})) \quad (16)$$

where $\sigma(\cdot)$ denotes the sigmoid function.

This formulation preserves the original CenterNet detection principle. Only the input representation of the heatmap head is refined. The size and offset branches remain unchanged.

Fig. 2 details the internal structure of FRM-hm, including the swSE-Block channel recalibration, the dilated spatial attention, the spatial gating mechanism, and the residual refinement block.

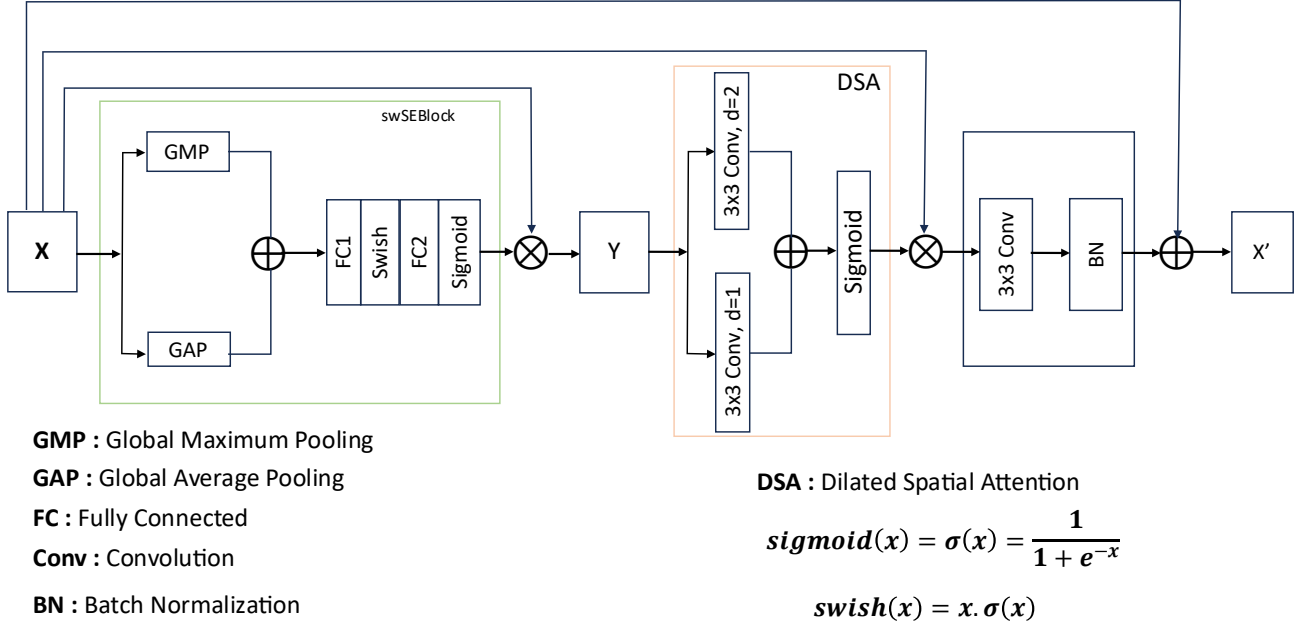


Fig. 2. Detailed architecture of the FRM-hm module.

3) Channel recalibration

The first component of FRM-hm is a channel recalibration mechanism inspired by squeeze-and-excitation strategies [17]. Its purpose is to emphasize channels that are more informative for object-center localization.

Given

$$X_{hm} \in \mathbb{R}^{C_h \times \frac{W}{r} \times \frac{H}{r}} \quad (17)$$

two global descriptors are first computed:

$$z_{avg} = \text{GAP}(X_{hm}), z_{max} = \text{GMP}(X_{hm}) \quad (18)$$

They are combined as

$$z = z_{avg} + z_{max} \quad (19)$$

The channel weights are then obtained through two fully connected layers separated by a Swish activation:

$$a = \text{FC}_2(\text{Swish}(\text{FC}_1(z))) \quad (20)$$

$$w_c = \sigma(a) \quad (21)$$

The recalibrated feature map is therefore

$$X_c = X_{hm} \odot w_c \quad (22)$$

where $\odot$ denotes the Hadamard product.

4) Dilated spatial attention and spatial gating

The second component of FRM-hm is a spatial attention mechanism dedicated to the localization of object centers. Starting from X_c , a spatial attention map is computed as

$$M_s = \sigma(\text{Conv}_{3 \times 3}^{d=1}(X_c) + \text{Conv}_{3 \times 3}^{d=2}(X_c)) \quad (23)$$

where the two convolutions use dilation rates $d = 1$ and $d = 2$, respectively. This design combines local detail and slightly larger contextual information while remaining lightweight.

The resulting attention map is then used to gate the original branch features:

$$X_g = X_{hm} \odot M_s \quad (24)$$

5) Residual refinement and output interpretation

A refinement block composed of a 3×3 convolution and batch normalization is then applied:

$$X_{res} = \text{BN}(\text{Conv}_{3 \times 3}(X_g)) \quad (25)$$

The final output of FRM-hm is obtained through a residual connection:

$$\tilde{X}_{hm} = X_{res} + X_{hm} \quad (26)$$

No additional nonlinear activation is applied after this final step. First, it preserves signed residual corrections, including negative responses that may suppress background clutter. Second, it avoids flattening subtle local feature variations that must later be converted into sharp and stable heatmap peaks. Third, the probabilistic constraint is already imposed by the sigmoid at the end of the heatmap head. Adding another non-linearity inside FRM-hm would therefore be redundant and could attenuate fine corrective information. This interpretation is consistent with the design logic already stated in the manuscript and with the implementation of the module.

Overall, REBDetX01 preserves the original CenterNet formulation while introducing a lightweight and localized refinement restricted to the most localization-sensitive branch of the detector. The size and offset heads remain unchanged, whereas the heatmap path benefits from a targeted feature refinement stage designed for the detection of small and visually challenging bird objects.

IV. RESULTS AND DISCUSSION

A. Dataset and Experimental Protocol

The performance of the proposed model REBDetX01 is evaluated on the FBD-SV-2024 dataset [13], which was introduced for the study of automatic bird detection under real surveillance conditions. The dataset contains 28,694 images with a resolution of 1280×720 pixels, extracted from videos recorded in an agricultural field.

FBD-SV-2024 focuses on the detection of perched and flying birds, which appear predominantly as small objects, with sizes ranging approximately from 10×10 to 70×70 pixels. Shape variability caused by motion, the weak visibility of certain birds, and the frequent presence of partial occlusions make the detection task particularly challenging.

These difficulties are further amplified by heterogeneous natural backgrounds, including fields, vegetation, and sky regions, as well as by variations in scale, viewpoint, and natural lighting conditions.

In this work, the task is formulated as a single-class discrete detection problem, where all birds are grouped into a single object category in accordance with the target application. Although the dataset originates from surveillance videos, only static frames are considered in this study, and detection is therefore performed in a frame-by-frame setting.

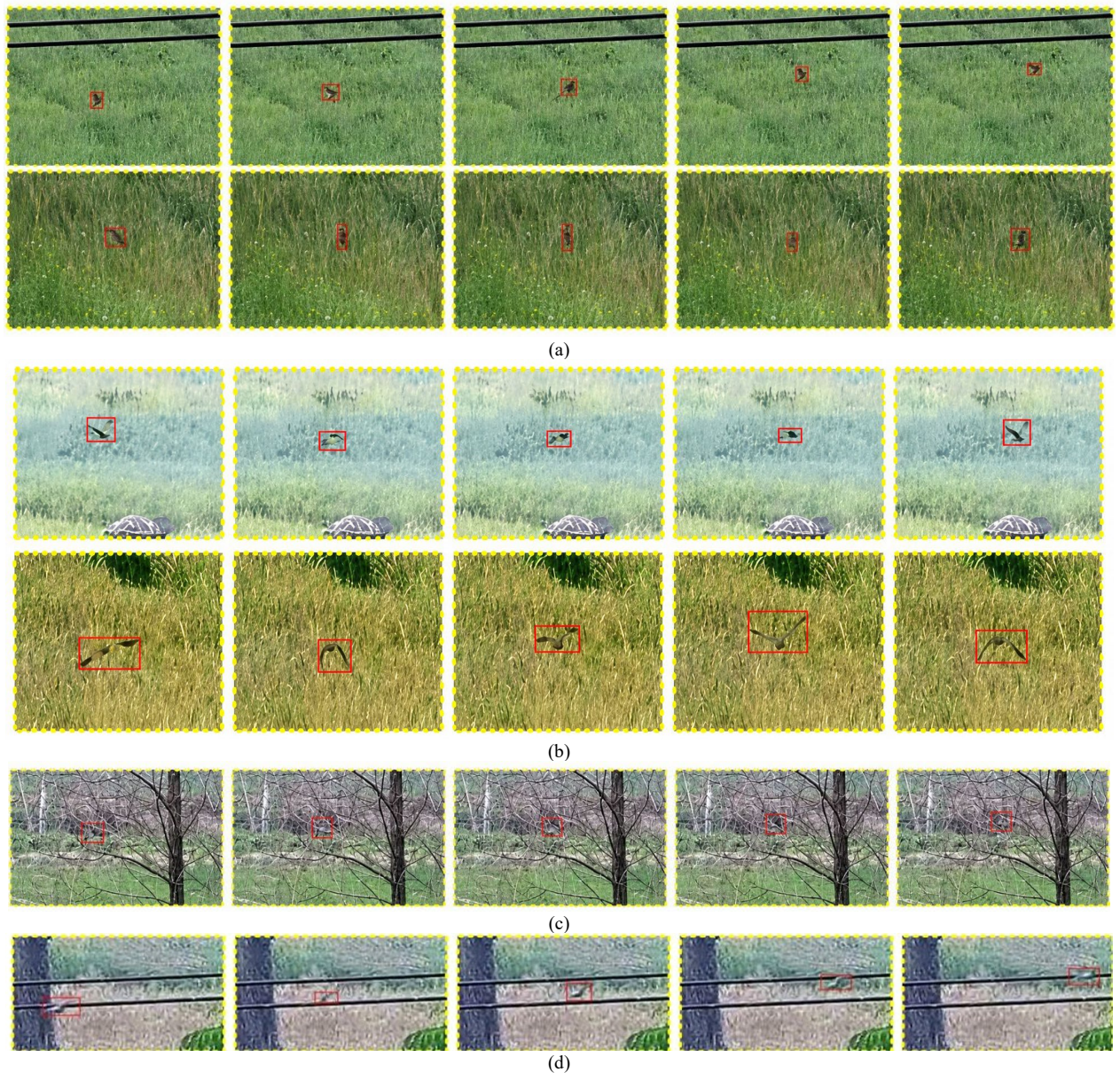


Fig. 3. Visual challenges in the FBD-SV-2024 dataset: (a) Low visibility of bird instances; (b) Motion-Induced shape variability; (c) Heterogeneous natural backgrounds; (d) Partial occlusions.

The experiments use the FBD-SV-2024 dataset, which contains 28,694 images with an original resolution of 1280×720 pixels. Following the official split, 23,979 images (83.57%) are used for training and 4715 images (16.43%) are used for testing, with no separate validation set defined in the original dataset. Among all images, 23,833 contain at least one bird instance, including 19,941 training images and 3892 test images. The remaining images do not contain birds and are treated as background frames. In total, the dataset includes 28,366 annotated bird instances. In this work, the task is addressed as a single-class frame-by-frame object detection problem.

The experiments follow the official FBD-SV-2024 train/test split. Therefore, the results reported in this work should be interpreted within this official train/test setting. The absence of an independent validation subset limits the possibility of performing rigorous hyperparameter tuning or checkpoint selection without increasing the risk of test-set overfitting.

The experiments in this study are conducted on FBD-SV-2024, which is selected because it is closely aligned with the target application considered in this work: the detection of small birds in complex natural surveillance scenes. Consequently, the quantitative conclusions reported in this paper are restricted to this dataset and to the described frame-by-frame single-class detection protocol. The present evaluation should therefore be interpreted as an in-dataset assessment of the proposed heatmap refinement strategy, rather than as a complete demonstration of cross-dataset generalization.

Fig. 3 highlights the main visual challenges of the FBD-SV-2024 dataset, including shape variability, low visibility, partial occlusions, complex natural backgrounds, and variations in scale, viewpoint, and illumination.

B. Training Settings and Implementation Details

The proposed REBDetX01 model is implemented in PyTorch and uses CenterNet as the baseline detection framework. The backbone is a ResNet-18 [32] network initialized with ImageNet pre-trained weights. In the implemented architecture, the standard convolutional stem and residual stages of ResNet-18 are retained as the feature extractor. A dropout layer with a rate of 0.5 is applied immediately after the backbone in order to improve regularization. The resulting features are then processed by a three-stage upsampling path that progressively increases the spatial resolution from stride 32 to stride 4, following the CenterNet design principle. As a result, for an input image resized to 512×512, the detector produces an output heatmap of size 128×128, corresponding to a reduction factor of 4.

The detection task is formulated as a single-class problem, and the number of object classes is therefore set to 1. On top of the shared upsampled feature map, three prediction heads are defined: a heatmap head, a size head, and an offset head. In accordance with the proposed architecture, the FRM-hm module is inserted only in the heatmap branch, while the size and offset branches remain unchanged. Each head follows the standard CenterNet design based on a 3×3 convolution, a ReLU activation, and

a final 1×1 convolution. In addition, the bias of the final classification layer in the heatmap head is initialized to -2.19, in order to compensate for the strong imbalance between the sparse object centers and the dominant background during the early stages of training.

Training is performed with a batch size of 8 using the Adam optimizer. The input resolution is fixed to 512×512. The global loss follows the CenterNet formulation and combines a focal loss for heatmap prediction with regression losses for object size and local offset. In the present implementation, the loss weights are set to 1.0 for the heatmap term, 1.0 for the offset term, and 0.1 for the size term. These settings are kept identical for REBDetX01, the baseline CenterNet model, and the ablation variants in order to ensure a fair comparison.

The optimization procedure uses a fixed two-stage training schedule applied identically to the baseline CenterNet model, REBDetX01, and all ablation variants. After one epoch with the pretrained ResNet-18 backbone frozen, the full network is fine-tuned for 250 epochs. At the unfreezing stage, the learning rate is reduced from 1×10^{-4} to 5×10^{-5} to limit abrupt changes in the pretrained backbone parameters. This choice provides a controlled comparison across model variants, but it is not claimed to be an optimal schedule. Alternative strategies such as cosine annealing or step-decay learning-rate schedules during joint fine-tuning may further improve adaptation to the small-object characteristics of FBD-SV-2024. As a result, each model is trained for a total of 251 epochs.

Training and evaluation are conducted on Google Colab Pro+, using an NVIDIA A100 GPU with 40 GB of VRAM, a dual-core Intel Xeon CPU at 2.20 GHz, 83.5 GB of RAM, and 235.7 GB of temporary disk storage. The software environment is based on Python 3, PyTorch, and CUDA 11.8. During evaluation, computational efficiency is measured on GPU in terms of Frames Per Second (FPS).

C. Evaluation Metrics

Model performance is evaluated using standard object detection metrics adapted to the present single-class discrete detection setting and to the small-object characteristics of the FBD-SV-2024 dataset. These metrics provide complementary information on localization accuracy, detection quality, robustness on small objects, and computational efficiency.

The primary metric used in this study is the Average Precision (AP), which summarizes the precision-recall trade-off over the set of detections. Let precision and recall be defined as

$$Precision = \frac{TP}{TP+FP} \quad (27)$$

$$Recall = \frac{TP}{TP+FN} \quad (28)$$

where TP , FP , and FN denote the numbers of true positives, false positives, and false negatives, respectively.

In the present single-class setting, AP is computed as the area under the precision-recall curve:

$$AP = \int_0^1 \text{Precision}(r) dr \quad (29)$$

where r denotes recall. In practice, AP is approximated by a discrete summation over the precision-recall curve:

$$AP = \sum_{n=1}^N (R_n - R_{n-1}) P_n \quad (30)$$

where N denotes the total number of sampled operating points on the precision-recall curve, R_n is the recall at the n -th point, R_{n-1} is the recall at the previous point, and P_n is the corresponding precision.

In addition to AP , two complementary AP variants are reported. The first is $AP50$, which corresponds to Average Precision computed at an Intersection over Union (IoU) threshold of 0.5. In the context of this study, $AP50$ provides a less restrictive measure of detection quality and is useful for assessing whether the model is able to detect bird instances even when the predicted boxes are not perfectly aligned with the ground truth. The second is APS , which measures Average Precision specifically for small objects. This metric is particularly important in the present work because FBD-SV-2024 is dominated by small bird instances, and the proposed FRM-hm module is explicitly designed to improve center localization under such conditions.

Inference efficiency is evaluated in terms of Frames Per Second (FPS) measured during Graphics Processing Unit (GPU) inference. This metric provides complementary information on the trade-off between detection accuracy and inference speed under the hardware setting used in this study. However, the reported FPS value should not be interpreted as a direct measure of embedded-device performance, since practical deployment efficiency also depends on runtime backend, operator implementation, memory access, model export strategy, and hardware-specific optimization.

D. Quantitative Results and Comparison

This subsection presents the quantitative results obtained by the proposed model REBDetX01 and by the reference models evaluated on the FBD-SV-2024 test set. Performance is analyzed using the metrics defined in Section IV-C, with particular emphasis on the detection of small objects, which dominate this dataset.

Table I reports the performance of the baseline CenterNet model as well as the variants obtained through the ablation study, where the different components of the FRM-hm module are introduced individually or jointly. The full REBDetX01 model, which combines swSE-Block-based channel recalibration and dilated spatial attention in the heatmap prediction branch, achieves the best overall results in terms of AP , $AP50$, and APS .

TABLE I. QUANTITATIVE RESULTS OF REBDetX01 AND ITS VARIANTS EVALUATED ON FBD-SV-2024

Model	AP (%)	AP50 (%)	APS (%)
CenterNet (baseline)	63.36	86.07	61.84
CenterNet + swSE-Block	63.05	86.37	61.04
CenterNet + DSA	63.11	86.29	61.38
REBDetX01	65.52	87.90	63.03

Compared with the baseline CenterNet model, REBDetX01 improves AP from 63.36% to 65.52% (+2.16 points), $AP50$ from 86.07% to 87.90% (+1.83 points), and APS from 61.84% to 63.03% (+1.19 points). These results indicate that the proposed heatmap refinement module improves both overall detection performance and small-object detection performance on FBD-SV-2024. The interpretation is therefore based on AP , $AP50$, and APS only.

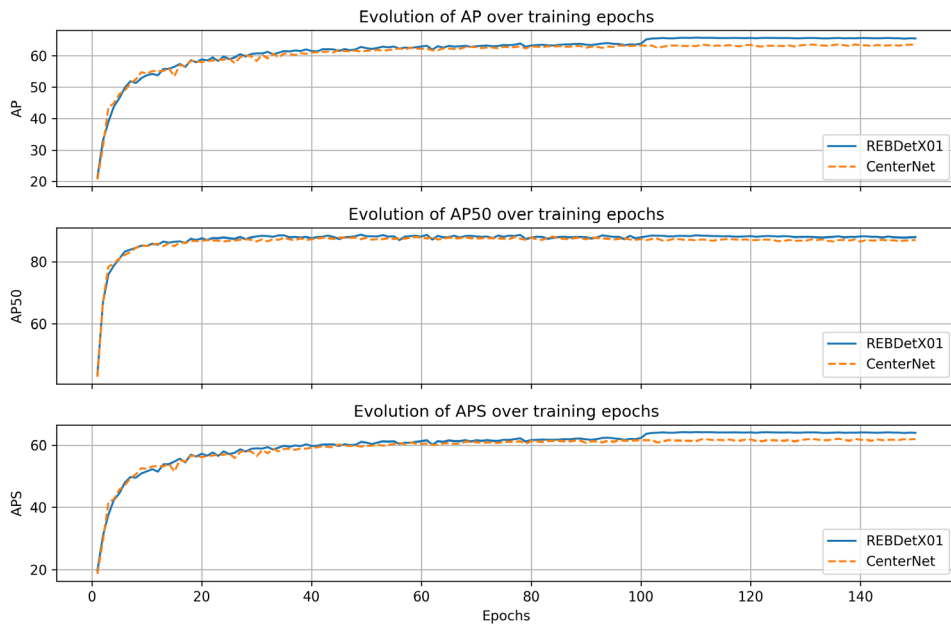


Fig. 4. Evolution of AP , $AP50$, and APS over the first 150 training epochs for REBDetX01 and CenterNet on the FBD-SV-2024 test set.

To analyze the evolution of performance during training, Fig. 4 presents the comparative curves of AP, AP50, and APS over the first 150 epochs for REBDetX01 and the baseline CenterNet model. These curves show that REBDetX01 reaches superior performance from the early stages of training and maintains a stable advantage in APS. This behavior confirms that the gains reported in Table I do not result from a late-stage training fluctuation, but rather reflect a structural improvement in the detection of small objects.

E. Computational Complexity Analysis

In addition to detection accuracy, the computational impact of the proposed module is an important consideration, especially for practical deployment. Based on the implementation of REBDetX01 and FRM-hm provided in the source code, the FRM-hm module operates on a 64-channel feature map after upsampling and contains three main components:

- A channel attention block based on Global Average Pooling (GAP), Global Max Pooling (GMP), and two fully connected layers,
- A dilated spatial attention block using two 3×3 convolutions with dilation rates $d = 1$ and $d = 2$,
- A residual refinement block composed of one 3×3 convolution followed by batch normalization.

The total number of trainable parameters introduced by FRM-hm is 38,724, which represents less than 0.5% of the parameters of the ResNet-18-based detector. For an input image of 512×512 pixels, corresponding to a heatmap branch feature map of size 128×128 , the module adds approximately 0.62 GFLOPs, which corresponds to about 1–2% of the total computational cost of the model.

This indicates that the proposed refinement strategy introduces only limited architectural and computational overhead. In this work, the term lightweight refers to this limited increase in parameters and FLOPs, rather than demonstrated optimization for embedded deployment.

Although FRM-hm introduces only a limited increase in the number of parameters and FLOPs, the FPS measured on GPU decreases from 1225.68 for the CenterNet baseline to 997.95 for REBDetX01. This observation indicates that practical inference speed is not determined only by parameter count or FLOP estimates. Operations such as dilated convolution, feature-map gating, and memory access patterns may affect runtime efficiency depending on the hardware backend and software implementation.

Compared with CBAM and SE-Block inserted at the same position, FRM-hm achieves higher detection performance under the same experimental protocol. However, this accuracy advantage should be considered together with runtime behavior. A complete latency-vs-accuracy analysis on embedded platforms would be necessary to determine whether the proposed module is preferable for deployment on resource-constrained devices.

Table II summarizes the computational complexity comparison between the baseline CenterNet model and the proposed REBDetX01 model.

TABLE II. COMPUTATIONAL COMPLEXITY COMPARISON BETWEEN CENTERNET AND REBDetX01

Model	Backbone	Parameters (M)	FLOPs (G)	FPS
CenterNet	ResNet-18	11.70	35.00	1225.68
REBDetX01	ResNet-18	11.74	35.62	997.95

F. Ablation Study

1) Internal ablation of FRM-hm

To analyze the relative contribution of each component of the proposed FRM-hm module, an ablation study was conducted on the FBD-SV-2024 dataset. The objective was to isolate the effect of channel recalibration and dilated spatial attention, and to evaluate their joint contribution when integrated into the heatmap prediction branch.

The results indicate that introducing either channel recalibration alone or dilated spatial attention alone does not systematically improve performance over the baseline CenterNet model. In some cases, these variants even lead to slightly lower results, especially on the APS metric. This observation highlights the difficulty of improving small-object detection by relying on a single local attention mechanism.

By contrast, when the two mechanisms are combined within FRM-hm, a consistent improvement is observed across all the considered metrics, with a particularly clear gain on APS. This result suggests that the two components play complementary roles. Channel recalibration improves the selection of informative feature channels, whereas dilated spatial attention enhances the spatially coherent regions associated with object centers.

Overall, these results show that the improvement brought by REBDetX01 does not arise merely from increased architectural complexity, but from the interaction of complementary mechanisms specifically designed for heatmap refinement.

2) Qualitative heatmap analysis of ablation variants

To further interpret the ablation results, a qualitative heatmap analysis was conducted on 100 images randomly selected from the dataset. For each image, the heatmap outputs of four variants were compared: CenterNet baseline, CenterNet + swSE-Block only, CenterNet + DSA only, and the complete REBDetX01 model. Fig. 5 presents six representative examples illustrating the general behavior observed across these visualizations.

The qualitative comparison shows that the baseline CenterNet model often produces heatmap responses around bird regions, but these activations may remain diffuse or affected by background structures. The swSE-Block-only variant improves the selection of informative feature channels, but this channel-level recalibration alone does not always produce a compact spatial response around object centers. The dilated spatial attention-only variant can reinforce local spatial responses, but without prior channel recalibration it may also amplify background-related activations or remain sensitive to local noise.

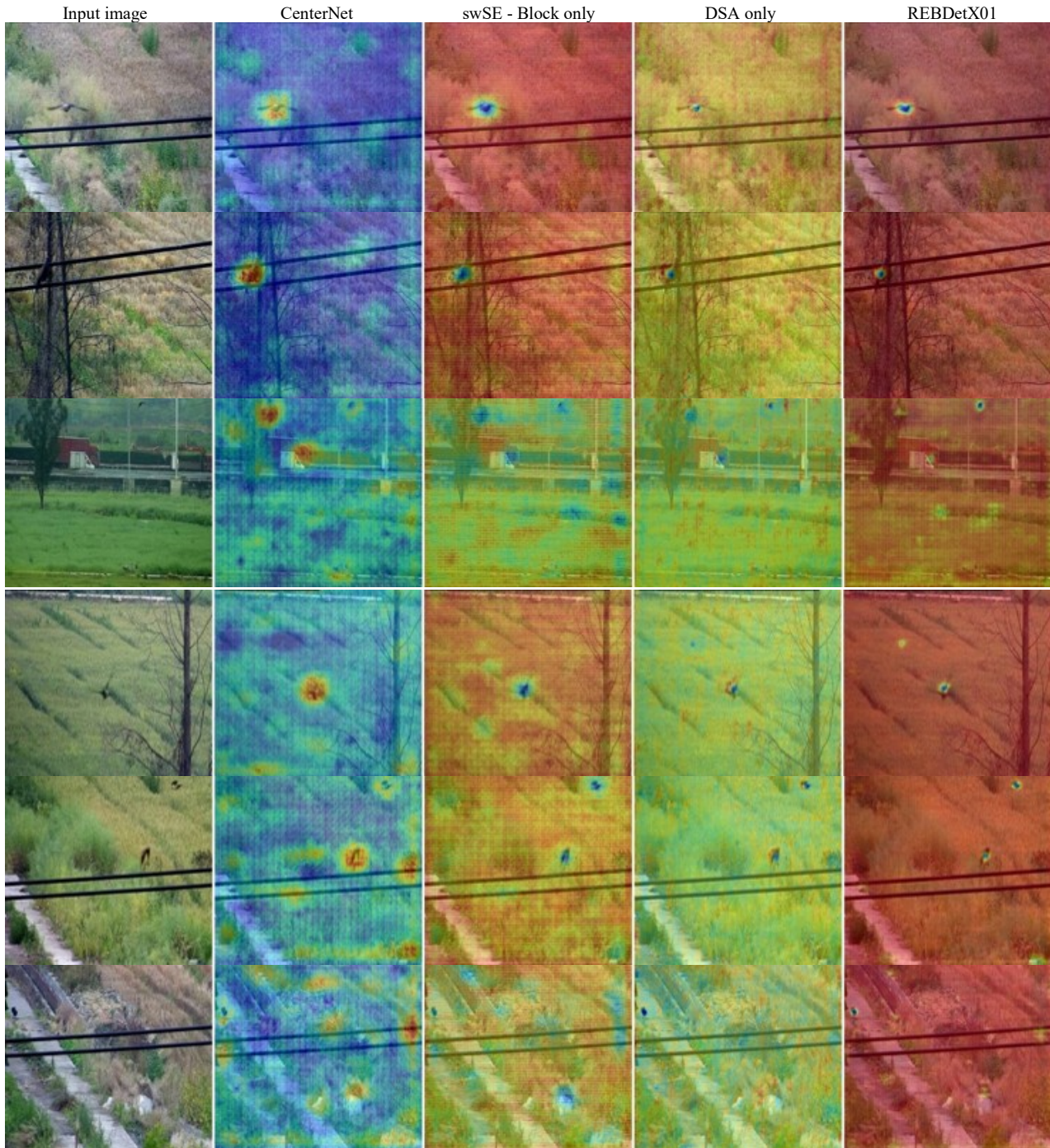


Fig. 5. Qualitative heatmap comparison of ablation variants on representative FBD-SV-2024 images.

By contrast, the complete FRM-hm module produces heatmaps that are generally more compact, better centered, and more spatially coherent around bird locations. This behavior suggests that the two components play complementary roles: swSE-Block improves feature selectivity at the channel level, while dilated spatial attention refines the spatial organization of responses associated with object centers. The residual refinement preserves local corrective information before the final heatmap prediction.

These observations help explain why the isolated components do not systematically improve the quantitative metrics, whereas their joint use within FRM-hm leads to

improved AP, AP50, and APS. The gain obtained by REBDetX01 is therefore not interpreted as a simple effect of adding attention, but as the result of a targeted coupling between channel recalibration and spatial refinement within the heatmap branch.

3) Comparison with alternative attention modules

To complement the internal ablation study, an additional comparison was conducted with two standard attention mechanisms integrated at the same architectural position as FRM-hm. In these variants, the proposed module was replaced by CBAM and SE-Block, respectively, while the rest of the detection architecture,

the training settings, and the evaluation protocol were kept unchanged. This additional experiment makes it possible to determine whether the observed gain comes from the specific design of FRM-hm or simply from the insertion of a generic attention mechanism into the heatmap branch.

As shown in Table III, replacing FRM-hm with CBAM or SE-Block does not yield the same level of performance as the proposed model. Both comparative variants remain below REBDetX01 on all the main detection metrics. Compared with the CBAM-based variant, REBDetX01 improves AP from 63.05% to 65.52% (+2.47 points), AP50 from 86.23% to 87.90% (+1.67 point), and APS from 61.42% to 63.03% (+1.61 point). Similarly, compared with the SE-Block-based variant, REBDetX01 improves AP from 62.91% to 65.52% (+2.61 points), AP50 from 86.09% to 87.90% (+1.81 point), and APS from 61.33% to 63.03% (+1.70 point).

TABLE III. COMPARATIVE EVALUATION OF FRM-HM AGAINST CBAM AND SE-BLOCK WHEN INSERTED INTO THE HEATMAP BRANCH UNDER IDENTICAL EXPERIMENTAL CONDITIONS

Model	AP (%)	AP50 (%)	APS (%)
CenterNet + CBAM	63.05	86.23	61.42
CenterNet + SE-Block	62.91	86.09	61.33
CenterNet + FRM-hm	65.52	87.90	63.03

This comparison is important for two reasons. First, it shows that the effectiveness of REBDetX01 cannot be attributed simply to the addition of a generic attention mechanism. If this were the case, CBAM and SE-Block would be expected to produce comparable gains. Second, the superiority of FRM-hm on APs is particularly meaningful in the present study, since the target dataset is dominated by small bird instances. This suggests that the

proposed module is better suited to refining the feature representation used for center localization under challenging small-object conditions.

It is also worth noting that the CBAM-based and SE-Block-based variants do not improve upon the original CenterNet baseline to the same extent as REBDetX01. This further supports the conclusion that the gain obtained with the proposed model does not result from attention insertion alone, but from the targeted design of FRM-hm as a dedicated heatmap refinement module.

Taken together, the results of the ablation study confirm that the effectiveness of REBDetX01 arises not from the generic use of attention, but from the specific combination of channel recalibration and dilated spatial attention within a module explicitly designed for the heatmap branch.

G. Qualitative Analysis

This subsection provides a qualitative analysis of the detection results obtained by REBDetX01 in comparison with the baseline CenterNet model, in order to better understand the impact of the FRM-hm module in complex scenes. The analysis is based on the visual inspection of detections produced on images from the FBD-SV-2024 test set. Fig. 6 provides a qualitative comparison of the heatmaps and detection results produced by the baseline CenterNet model and the proposed REBDetX01 model, respectively.

In Fig. 6(a), the baseline CenterNet result shows that bird centers can be detected, but the predicted heatmap responses remain relatively weak and less concentrated than the ground-truth peaks. This behavior may make center localization more sensitive to background clutter, weak contrast, and the small apparent size of the birds.

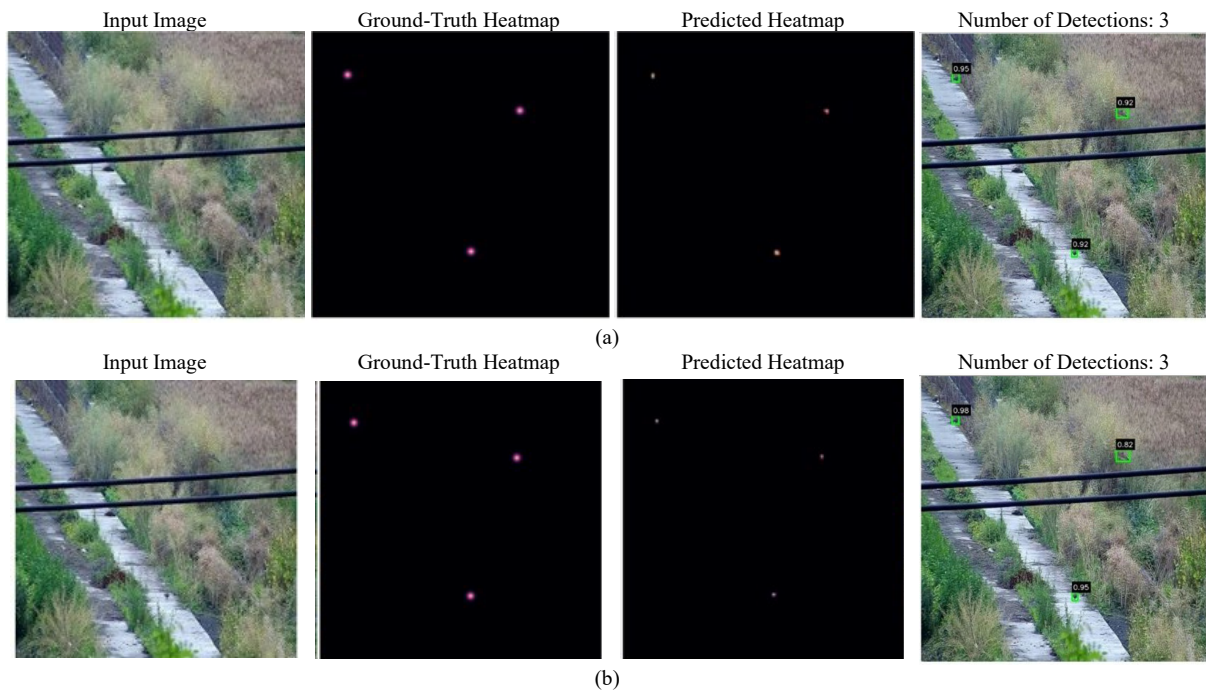


Fig. 6. Qualitative comparison of detection results on FBD-SV-2024 test images: (a) Prediction and ground truth with CenterNet; (b) Prediction and ground truth with REBDetX01.

In Fig. 6(b), the REBDetX01 result shows more compact and better localized heatmap responses around the bird centers. The detection result is therefore more consistent with the expected object-center positions, indicating that the FRM-hm module helps strengthen the local maxima used by the center-based detector.

The point-by-point comparison of Fig. 6(a) and Fig. 6(b) supports the quantitative findings: compared with the baseline, REBDetX01 produces clearer center responses and more stable detections for small birds in complex scenes. Nevertheless, some limitations remain. In scenes with high bird density, severe occlusions, or extremely low contrast between the object and the background, the model may still miss certain detections or produce imprecise localization. These cases illustrate the intrinsic difficulty of single-image bird detection in natural environments.

H. Comparison with Reference Methods on FBD-SV-2024

To evaluate the positioning of the proposed model with respect to existing work, the performance of REBDetX01 is compared with the results reported in Table III of Sun *et al.* [13] for the same dataset.

However, comparisons across studies must be interpreted with caution because they require strict methodological alignment, especially regarding evaluation metrics, training protocols, and experimental settings.

To position REBDetX01 with respect to existing methods, we compare it with the results reported by Sun *et al.* [13] on FBD-SV-2024 for recent image-based and video-based detectors, including YOLOv5l, YOLOv6l, YOLOXl, YOLOv8l, YOLOv9e, YOLOv10l, SSD, FBOD-BMI, and FBOD-SV. These methods represent different detection paradigms, including frame-by-frame image detection and spatio-temporal video-based detection. However, the comparison must be interpreted as contextual rather than strictly equivalent, since differences in training protocol, input resolution, initialization, evaluation scripts, and post-processing settings may affect the reported metrics. Within this comparative framework, REBDetX01 is positioned as a heatmap-refinement approach designed to improve center-based small-object detection without modifying the backbone architecture or exploiting temporal information.

1) Methodological alignment and comparative framework

Sun *et al.* [13] state that AP is measured according to the Pascal Visual Object Classes (VOC) 2007 protocol [21]. However, their Table III simultaneously reports metrics such as AP, AP50, AP75, APS, Average Precision for medium objects (APM), and Average Precision for large objects (APL), which are typically associated with the Microsoft COCO evaluation protocol [20].

These two protocols differ in several aspects:

- interpolation of Precision-Recall curves,
- number of IoU thresholds used,
- object size categorization,
- final score aggregation.

In the absence of an explicit correspondence between the evaluation scripts, the numerical values obtained under the two protocols cannot be considered strictly equivalent. Therefore, the comparison provided here should be regarded as indicative rather than absolute.

In addition, Sun *et al.* [13] specify that all models in their study are trained from scratch, without external pretraining. This choice may affect the performance of detectors such as YOLO, which are often trained under different conditions in standard benchmarks.

Possible differences in input image size, resizing strategy, score thresholds, IoU assignment rules, and non-maximum suppression settings may also contribute to the observed performance differences.

Several external datasets could be considered for future cross-dataset evaluation, including bird-detection datasets collected around wind farms, such as Klim and Skagen [33, 34], as well as recent datasets addressing distant or small bird detection [22–24].

However, these datasets differ from FBD-SV-2024 in important aspects, including image modality, acquisition environment, background structure, object scale distribution, preprocessing strategy, temporal usage, and evaluation protocol. For instance, wind-farm datasets are often composed of grayscale sky-oriented images and may rely on tiling or temporal stacking to improve small-object detection. Therefore, a direct numerical comparison without harmonizing the training and evaluation conditions could lead to misleading conclusions. A fair comparison would require retraining and evaluating both the CenterNet baseline and REBDetX01 under the same controlled protocol on each external dataset.

2) Comparison with frame-by-frame image detection models

Among the methods reported in Sun *et al.* [13], the YOLO family constitutes relevant reference models in the frame-by-frame detection setting.

These detectors rely on an anchor-based paradigm, in which classes and bounding boxes are predicted simultaneously from dense grids. They are generally designed to balance detection accuracy and inference speed.

In contrast, REBDetX01 follows an anchor-free center-based paradigm, in which detection is driven by the prediction of object centers, with a targeted refinement of the heatmap branch through FRM-hm.

The differences observed between the reported YOLO results and those obtained by REBDetX01 cannot be attributed solely to architectural differences, since protocol and training differences also play an important role. However, from a structural point of view, center-based detection offers a potential advantage for the localization of very small objects, as it avoids dependence on predefined anchor sets whose configuration may become critical in dense scenes containing tiny objects.

The improvements observed on AP_s confirm that FRM-hm strengthens the consistency and saliency of local heatmap maxima, thereby improving the detection of small birds in the FBD-SV-2024 dataset.

3) Analysis of models explicitly exploiting temporal information

The models FBOD-BMI and FBOD-SV proposed by Sun *et al.* [13] belong to a different paradigm, as they explicitly exploit temporal information from image sequences.

FBOD-SV integrates a temporal aggregation mechanism based on Co-Attention Feature Aggregation, which combines information from multiple successive frames before or during feature extraction. It also relies on a dedicated architecture, FBOD-Net, combining large receptive field layers and high-resolution layers to improve multi-scale object detection.

FBOD-BMI follows a motion-centered approach. It incorporates a ConvLSTM-PAN module for spatio-temporal aggregation, a tracking mechanism to estimate the motion range, and the generation of adaptive spatio-temporal cubes to refine final detection.

These architectures have a structural advantage in situations where an object may be blurred, partially occluded, or weakly contrasted in one frame but more

visible in adjacent frames. Inter-frame coherence can thus reinforce weak features and improve robustness.

By contrast, REBDetX01 operates strictly in a single-image, frame-by-frame setting, without exploiting temporal cues. Therefore, numerical comparisons between these paradigms must be interpreted with this fundamental difference in mind.

4) Overall positioning of REBDetX01

Taking the above considerations into account, Table IV shows that REBDetX01 is positioned favorably among frame-by-frame detection approaches evaluated on FBD-SV-2024, particularly for the detection of small objects.

Importantly, the gains achieved by REBDetX01 result from a targeted modification of the heatmap branch through FRM-hm, without altering the backbone

architecture and without exploiting temporal information. This supports the idea that localized refinement of heatmap features constitutes an effective strategy for improving center-based detection in complex natural scenes.

TABLE IV. COMPARISON OF REBDetX01 WITH REFERENCE METHODS EVALUATED ON THE FBD-SV-2024 DATASET

Model	Setting	Paradigm	Image Size	AP (%)	AP50 (%)	APS (%)
YOLOv5l	frame-by-frame	anchor-based	640×640	29.9	55.8	22.0
YOLOv6l	frame-by-frame	anchor-based	640×640	30.4	58.5	21.4
YOLOXl	frame-by-frame	anchor-based	640×640	33.8	62.0	28.0
YOLOv8l	frame-by-frame	anchor-based	640×640	31.8	58.4	23.3
YOLOv9e	frame-by-frame	anchor-based	640×640	31.8	57.7	22.7
YOLOv10l	frame-by-frame	anchor-based	640×640	29.4	55.0	19.9
SSD	frame-by-frame	anchor-based	640×640	29.9	59.9	23.1
CenterNet (baseline)	frame-by-frame	anchor-free / center-based	512×512	63.36	86.07	61.84
CenterNet + swSE-Block	frame-by-frame	anchor-free / center-based	512×512	63.05	86.37	61.04
CenterNet + DSA	frame-by-frame	anchor-free / center-based	512×512	63.11	86.29	61.38
REBDetX01	frame-by-frame	anchor-free / center-based	512×512	65.52	87.90	63.03
FBOD-BMI	image sequence	spatio-temporal	672×384	35.1	69.2	30.6
FBOD-SV	image sequence	spatio-temporal	672×384	37.1	71.9	31.3

V. CONCLUSION AND FUTURE WORKS

In this work, we proposed FRM-hm, a dedicated heatmap refinement module integrated into a CenterNet-based object detector, resulting in the model REBDetX01, for improving the detection of small granivorous birds in rice-field images. The proposed module combines channel recalibration and dilated spatial attention to enhance the localization of object centers in complex natural scenes, while preserving the original backbone architecture of the detector. Overall, the main strength of REBDetX01 is that it improves center-based small-object detection through a localized and interpretable heatmap refinement strategy.

Experimental results on the FBD-SV-2024 dataset demonstrate that REBDetX01 improves detection performance compared with the baseline CenterNet model. In particular, the proposed approach achieves gains in AP, AP50, and APS, indicating enhanced capability for detecting small objects. The ablation study further confirms that these improvements arise from the complementary interaction between the channel recalibration mechanism and the dilated spatial attention

integrated within the FRM-hm module. The quantitative results, ablation study, and comparison with generic attention modules show that this targeted design improves the detection of small birds in complex natural scenes with only a limited additional computational cost.

Despite these encouraging results, the present study has several limitations. First, the evaluation is restricted to the FBD-SV-2024 dataset. Although this dataset is well aligned with the target application and contains challenging small bird instances, the current results do not fully demonstrate cross-dataset generalization. Second, REBDetX01 is formulated as a frame-by-frame detector and does not exploit temporal information available in surveillance videos, which may be useful under motion blur, weak visibility, or temporary occlusion. Third, although FRM-hm introduces only a limited increase in parameters and FLOPs, no dedicated evaluation has yet been conducted on embedded devices such as NVIDIA Jetson Nano. Fourth, the current task is formulated as single-class bird detection, whereas future practical applications may require multi-class or species-level detection. Finally, the present study follows the official FBD-SV-2024 train/test partition, which does not include a separate validation set. As a result, the current

experiments do not provide a dedicated validation subset for systematic hyperparameter tuning, checkpoint selection, or learning-rate schedule comparison.

Future work will address these limitations by conducting external-dataset validation, a video-level Train/Validation/Test split to evaluate optimization strategies more rigorously while avoiding possible frame-level leakage between subsets, temporal extension of the model, deployment-oriented benchmarking on embedded platforms, and extension to richer detection settings such as multi-class or species-level bird detection.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

G.L.L. provided overall scientific supervision, defined the research framework, and designed the study methodology. K.B.Y. carried out data preparation, data analysis, model development, and manuscript drafting. A.K.A. contributed to model assessment, result validation, and supervision of the manuscript writing process. All authors approved the final version.

REFERENCES

- [1] P. M. del Villar and F. Lançon, "West African rice development: Beyond protectionism versus liberalization?" *Global Food Security*, vol. 5, pp. 56–61, 2015.
- [2] G. Soullier, M. Demont, A. Arouna *et al.*, "The state of rice value chain upgrading in West Africa," *Global Food Security*, vol. 25, 100365, 2020.
- [3] F. Tondel, C. D'Alessandro, I. Hathie, and C. Blancher. (2020). Rice trade and rice sector development in West Africa: An approach for more coherent public policies. European Centre for Development Policy Management (ECDPM) and the Initiative Prospective Agricole et Rurale (IPAR). [Online]. Available: <https://ecdpm.org/work/rice-trade-and-value-chain-development-in-west-africa-an-approach-for-more-coherent-policies>
- [4] K. G. Odoukpe and K. H. Yaokoré-Béibro, "Evaluation of damage caused to rice fields by birds in the Grand-Bassam wetland (Côte d'Ivoire)," *African Journals Online (AJOL) (Agronomie Africaine)*, vol. 28, no. 3, pp. 42–52, 2016. (in French)
- [5] I. Yattara, S. Sissoko, M. S. Maïga *et al.*, *Influence of Granivorous Birds on Cereal Production in the Lake Debo and Lake Walado Debo Area, Lower Niger River Delta (French Edition)*, Éditions Universitaires Européennes, 2019, pp. 45–52. (in French)
- [6] P. Marañón *et al.*, "A system using artificial intelligence to detect and scare bird flocks in the protection of ripening fruit," *Sensors*, vol. 21, no. 12, 4244, 2021.
- [7] S. Manikowski, A. B. Ndiaye, and T. Bernard, "Manual for crop protection against bird damage," *Food and Agriculture Organization of the United Nations, Dakar, West Africa*, 1991. (in French)
- [8] A. Bouet, A. Boka, and N. Kouassi, "Impact of human monitoring on bird damage in rainfed rice cultivation," *International Journal of Biological and Chemical Sciences*, vol. 8, no. 5, pp. 2314–2319, 2014. (in French)
- [9] K. E. S. Kouadja, A. J. Gnago, K. S. G. Odoukpe *et al.*, "Bird diversity and impact of net use in controlling rice-depredating birds in a lowland rice field of Yamoussoukro (central Côte d'Ivoire)," *Agronomie Africaine*, vol. 33, no. 1, pp. 107–118, 2021. (in French)
- [10] A. Bouet *et al.*, "Effectiveness of a methyl anthranilate-based repellent on bird pressure in strict rainfed rice cultivation," *International Journal of Biological and Chemical Sciences*, vol. 11, no. 6, pp. 2856–2866, 2017. (in French)
- [11] R. Girshick, J. Donahue, T. Darrell *et al.*, "Rich Feature hierarchies for accurate object detection and semantic segmentation", in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.
- [12] J. Redmon, S. Divvala, R. Girshick *et al.*, "You only look once: Unified, real-time object detection," in *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91
- [13] Z. W. Sun *et al.*, "Fbd-sv-2024: Flying bird object detection dataset in surveillance video," *Scientific Data*, vol. 12, no. 1, 530, 2025.
- [14] T. Y. Lin, P. Dollar, R. Girshick *et al.*, "Feature pyramid networks for object detection," in *Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2117–2125. doi: 10.1109/CVPR.2017.106
- [15] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv preprint, arXiv:1804.02767, 2018.
- [16] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv Preprint*, arXiv:1904.07850, 2019.
- [17] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.
- [18] S. Woo, J. Park, J. Y. Lee *et al.*, "Cbam: Convolutional block attention module," in *Proc. European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19. doi: 10.1007/978-3-030-01234-2_1
- [19] B. Tréca, A. B. Ndiaye, and S. Manikowski, *Crop-Depredating Birds in the Sahel*, Paris, France: J. Libbey Eurotext, 1997. (in French).
- [20] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Proc. European Conference on Computer Vision*, 2014, pp. 740–755.
- [21] M. Everingham, L. V. Gool, C. K. I. Williams *et al.*, "The pascal Visual Object Classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010. doi: 10.1007/s11263-009-0275-4
- [22] S. Fujii, K. Akita, and N. Ukita, "Distant bird detection for safe drone flight and its dataset," in *Proc. 2021 17th International Conference on Machine Vision and Applications (MVA)*, 2021, pp. 1–5.
- [23] Y. Kondo *et al.*, "MVA2023 small object detection challenge for spotting birds: Dataset, methods, and results," in *Proc. 2023 18th International Conference on Machine Vision and Applications (MVA)*, 2023, pp. 1–11. doi: 10.23919/MVA57639.2023.10215935
- [24] H. Sun *et al.*, "AirBirds: A large-scale challenging dataset for bird strike prevention in real-world airports," in *Proc. Asian Conference on Computer Vision*, 2022, pp. 2440–2456. doi: 10.1007/978-3-031-26348-4_24
- [25] Z. W. Sun, Z. X. Hua, H. C. Li *et al.*, "Flying bird object detection algorithm in surveillance video based on motion information," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–15, 2023.
- [26] Z. Sun, Z. Hua *et al.*, "A flying bird object detection method for surveillance video," arXiv preprint, arXiv:2401.03749, 2024.
- [27] S. Ren, K. He, R. Girshick *et al.*, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.
- [28] Z. Ge, S. Liu, F. Wang *et al.*, "Yolox: Exceeding yolo series in 2021," arXiv preprint, arXiv:2107.08430, 2021.
- [29] C. Y. Wang, I. H. Yeh, and H. Y. M. Liao, "Yolov9: Learning what you want to learn using programmable gradient information," in *Proc. European Conference on Computer Vision*, 2024, pp. 1–21. doi: 10.1007/978-3-031-72751-1_1
- [30] A. Wang *et al.*, "Yolov10: Real-time end-to-end object detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 107984–108011, 2024.
- [31] Y. Zhao *et al.*, "Detrs beat yolos on real-time object detection," in *Proc. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 16965–16974. doi: 10.1109/CVPR52733.2024.01605
- [32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90
- [33] R. Yoshihashi, R. Kawakami, M. Iida *et al.*, "Bird detection and species classification with time-lapse images around a wind farm: Dataset construction and evaluation," *Wind Energy*, vol. 20, no. 12, pp. 1983–1995, 2017. doi: 10.1002/we.2135

- [34] H. Alqaysi, I. Fedorov, F. Z. Qureshi *et al.*, “A temporal boosted yolo-based model for birds detection around wind farms,” *Journal of Imaging*, vol. 7, no. 11, 227, 2021. doi: 10.3390/jimaging7110227

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited ([CC BY 4.0](#)).