Multi-Crop Reproductive Structure Detection and Counting Using a Lightweight YOLOv8n with Spatial–Channel Attention and Re-parameterizable Convolutions

Md. Mushibur Rahman MUSHIBURRAHMAN5@GMAIL.COM

Umme Fawzia Rahim FAWZIA.RAHIM@DUET.AC.BD

Department of Computer Science and Engineering
Dhaka University of Engineering & Technology
Gazipur, Bangladesh

Editor: Sakinat Folorunso, Roseline Ogundokun, and Francisca Oladipo

Abstract

Detecting and counting reproductive structures across multiple crop types in real agricultural environments remains challenging due to occlusion, large-scale variations, changing illumination, and inter-crop visual differences. Although deep learning-based detectors have advanced considerably, their performance often declines in densely populated and visually complex scenes, limiting practical deployment in resource-constrained agricultural settings. This study presents a lightweight YOLOv8n-based framework for multi-crop reproductive structure detection and counting. A Spatial–Channel Attention Convolution module integrated into cross-stage fusion (C2f-SCA) highlights informative spatial regions and feature channels, while Re-parameterizable Depthwise Convolution (RepDWConv) strengthens multi-scale feature representation without increasing inference cost. The framework is evaluated on three heterogeneous datasets: Cauliflower, Tomato Flower, and Maize Tassel. Experimental results show that the proposed model achieves mAP@0.5 of 0.981, 0.969, and 0.965 with F1-scores of 0.954, 0.921, and 0.936, respectively. Counting accuracy improves consistently across all datasets, with MAE of 0.258, 0.210, and 1.068, and R^2 of 0.965, 0.962, and 0.996 for Cauliflower, Tomato Flower, and Maize Tassel, respectively. These improvements are achieved with approximately 12% fewer parameters than the YOLOv8n baseline, demonstrating a compact and generalizable solution suitable for automated crop monitoring, yield estimation, and scalable precision agriculture in resource-constrained environments.

Keywords: Lightweight Object Detection, YOLOv8, Attention Mechanisms, Multi-Crop Plant Phenotyping, Object Counting, Precision Agriculture

1. Introduction

Accurate detection and counting of plant organs such as flowers, fruits, and tassels are essential tasks in plant phenotyping and precision agriculture **?**. However, reliable trait measurement in field environments remains challenging due to variations in lighting, occlusion, and large variations in object scale. Deep learning-based object detection, particularly single-stage YOLO models, has become widely adopted due to their strong balance between accuracy and real-time performance. Recent studies show that architectural improvements to YOLO significantly enhance plant reproductive structure detection performance. For

instance, **?** proposed an improved YOLOv8-based model for soybean pod detection, integrating a Diverse Branch Block (DBB) backbone and Spatially Enhanced Attention Module (SEAM) to enhance feature representation, achieving 83.1% AP on their dataset. Similarly, **?** proposed the YOLO-MCS model for loquat fruit detection in complex orchard environments, achieving 89.9% mAP while reducing model parameters and computational cost. Lightweight convolution and attention mechanisms have also demonstrated strong performance for maize tassel detection in UAV imagery, achieving mAP@0.5 above 91% while maintaining real-time performance **?**. Further improvements using global attention and multi-branch feature aggregation have reported mAP@0.5 exceeding 96% for maize tassel detection **?**. YOLO-based models have also been applied to tomato flower and fruit detection, achieving 95.3% mAP@0.5 using depthwise separable convolutions with Squeeze-and-Excitation (SE) attention **?**.

Despite these promising advancements, several critical gaps remain. Most existing approaches focus on single-crop scenarios and rely on proprietary datasets that limit reproducibility, and often employ complex attention architectures that increase computational cost. Consequently, only a few studies address detection and counting jointly within a unified framework. To address these limitations, this study proposes a lightweight YOLOv8-based framework for multi-crop reproductive structure detection and counting that integrates Spatial–Channel Attention (SCA) within the cross-stage fusion module (C2f-SCA) and re-parameterizable depthwise convolution (RepDWConv). These modules enhance feature discrimination and multi-scale fusion while maintaining low computational overhead, enabling robust generalization across multiple crop types and improved detection and counting accuracy. In light of these research gaps, the specific objectives delineated in this study are:

- We propose a lightweight YOLOv8n-based detection framework that incorporates C2f-SCA and RepDWConv to enhance multi-scale feature representation while maintaining computational efficiency.
- We introduce a real-world open-field cauliflower dataset collected and manually annotated under natural agricultural conditions to support research on plant reproductive structure detection and counting.
- Comprehensive experiments across three heterogeneous datasets (Cauliflower, Tomato Flower, and Maize Tassel) demonstrate that the proposed framework improves detection accuracy and counting reliability while maintaining a compact architecture suitable for real-time agricultural applications.

2. Related Work

Recent advances in deep learning have substantially improved automatic plant reproductive structure detection and counting, which play an important role in precision agriculture, crop monitoring, and yield estimation. Convolutional neural network-based detectors, particularly YOLO variants, have become widely adopted because they offer a favorable balance between detection accuracy and real-time inference speed. Earlier studies often relied on multi-stage detection frameworks. For example, Faster R-CNN demonstrated strong performance for seed and fruit counting in controlled phenotyping environments **?**. Similarly, the

Sugarcane-Detector (SGN-D) combined channel attention with multi-scale feature fusion to enhance UAV-based sugarcane seedling detection and counting **?**. Another research direction focuses on regression-based counting approaches that estimate object numbers directly without explicit localization. These methods can achieve competitive results while maintaining relatively compact model structures **?**. However, regression-based approaches often struggle in dense scenes and provide limited interpretability compared with detection-based methods. More recently, researchers have enhanced YOLO architectures using lightweight convolutions, attention modules, and transformer-based feature fusion, achieving AP values up to 94.99% in maize tassel detection tasks **??**.

YOLO-based architectures have also been widely used for plant reproductive structure detection across different crops. FlowerYOLOv5 improved YOLOv5s by integrating Convolutional Block Attention Module (CBAM) and enhanced feature fusion, achieving 94.2% on a tomato flower dataset **?**. Cauli-Det further modified YOLOv8s by adding convolutional blocks and replacing SiLU with Hard-Swish activation, improving mAP@0.5 to 91.1% on a cauliflower detection dataset **?**. Attention-based YOLOv8 variants have also been explored for tomato maturity grading and counting tasks, where MHSA-YOLOv8 improved contextual feature modeling for tomato maturity grading and counting on a tomato dataset, achieving competitive detection and counting performance **?**. Several studies integrate tracking algorithms such as YOLOv5 with DeepSort to improve counting reliability in UAV-based tomato greenhouse datasets, although performance decreased in dense scenes due to occlusion and class imbalance **?**. Other works combined YOLO with StrongSORT or ByteTrack for agricultural video datasets, including rapeseed flower and dragon fruit monitoring videos, achieving AP values above 94% and counting accuracy exceeding 94.51% **??**. Despite these advances, many existing approaches rely on private datasets, focus on single-crop scenarios, or depend on complex tracking modules. Therefore, developing lightweight detection frameworks that maintain high accuracy while supporting cross-crop generalization remains an important research direction.

3. Methodology

3.1. Datasets

To evaluate the robustness and cross-crop generalization of the proposed framework, experiments were conducted on three datasets: (1) a self-collected open-field cauliflower dataset, (2) a greenhouse tomato flower dataset **?**, and (3) a maize tassel dataset **?**. These datasets differ in crop type, imaging conditions, object density, regional differences, and scale, and include challenges such as leaf occlusion and object overlap, providing a diverse evaluation setting.

3.1.1. Cauliflower Dataset

This study introduces a new open-field cauliflower dataset collected from four agricultural sites across the Bogura and Dhaka districts of Bangladesh during the winter cultivation season (November 2025–January 2026). Images were captured when plants were 52 and 65 days old during the curd formation stage. Images were acquired using Google Pixel 7 and Redmi Note 10S smartphones. Data collection was performed under natural lighting

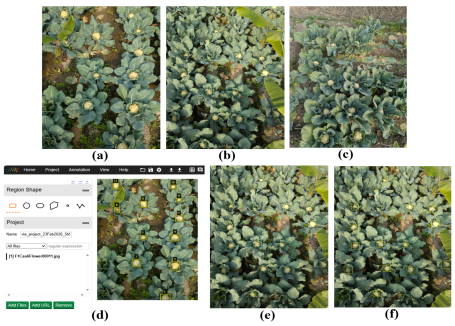

Figure 1: Sample cauliflower images captured under (a) morning (10:00 to 11:00 AM), (b) afternoon (12:30 to 1:30 PM), (c) evening (4:00 to 5:00 PM). (d) VIA annotation interface, (e) original image, and (f) annotated images.

conditions during morning, afternoon, and evening periods, with ambient temperatures ranging from 20–24 °C. The dataset includes images with resolutions of 3072 × 4080, 2464 × 3280, and 3472 × 4624 pixels. Overall, the dataset includes 7,683 annotated cauliflower instances, consisting of 6,144 training, 760 validation, and 779 test instances. All images were manually annotated using the VGG Image Annotator (VIA) and verified through manual inspection. Representative samples of the dataset along with their ground-truth annotations are shown in Figure 1.

3.1.2. Tomato Flower Dataset

The tomato flower dataset used in this study was originally introduced by **?**. The images were captured in a greenhouse at the Faculty of Agriculture, Shizuoka University, Japan, under natural daylight conditions. For this study, a subset of 2,000 images was selected and divided into 1,600 training, 200 validation, and 200 test images, with all images resized to 600 × 600 pixels. The dataset contains 8,262 annotated tomato flower instances, including 6,549 for training, 872 for validation, and 841 for testing.

3.1.3. Maize Tassel Dataset

The maize tassel dataset used in this study is derived from the Maize Tassel Detection and Counting (MTDC) dataset proposed by **?**. In total, 3,079 images were used, including 2,463 for training, 307 validation, and 309 testing. The dataset includes 126,582 annotated maize tassel instances, with 101,420 in training, 12,225 in validation, and 12,937 in the test set. Each tassel is annotated with a bounding box, and the task is treated as a single-class detection and counting.

3.2. YOLOv8n

The multi-crop plant reproductive structure detection and counting framework is built upon the YOLOv8n architecture, which is adopted as the baseline due to its anchor-free design

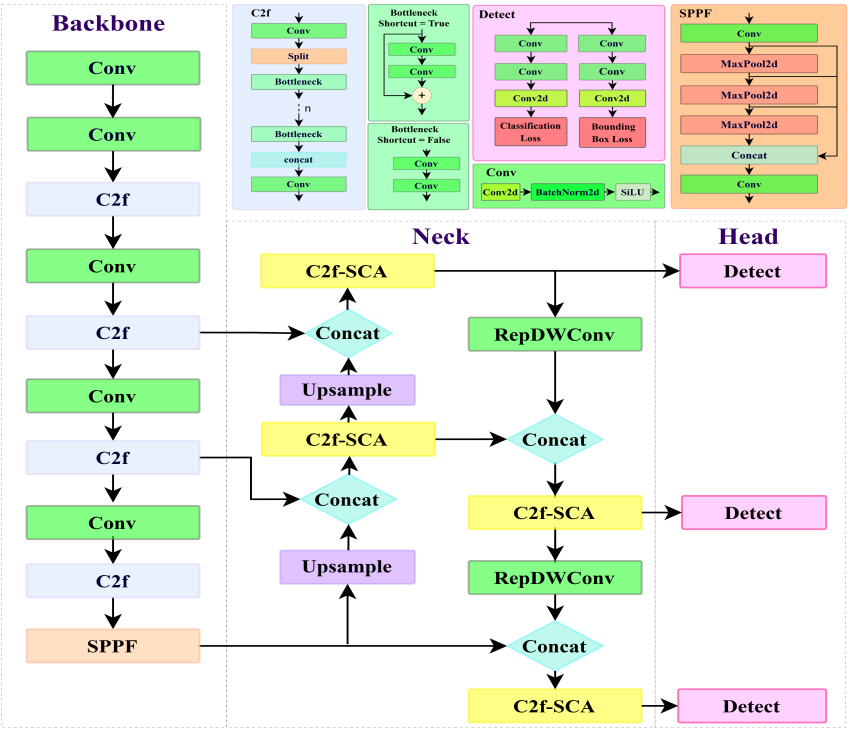

Figure 2: Overall architecture of the proposed lightweight YOLOv8

and strong balance between detection accuracy and computational efficiency **?**. YOLOv8 is a one-stage detector that combines lightweight convolution, hierarchical feature representation, and multi-scale fusion, making it suitable for agricultural scenes where objects are small, variable in scale, and often partially occluded. A key component of the backbone is the C2f (Cross-Stage Partial Fusion) module, which improves gradient flow and feature reuse through split-transform-concatenate operations. The backbone includes a Spatial Pyramid Pooling-Fast (SPPF) module that collects contextual information at multiple scales through repeated max-pooling operations, helping the model detect objects of different sizes in complex backgrounds **?**. The neck of YOLOv8 adopts a multi-scale feature fusion strategy inspired by the Feature Pyramid Network (FPN) **?**, where high-level semantic features are fused with lower-level spatial features through upsampling and concatenation. This design improves detection performance for objects with significant scale variation and is particularly beneficial for identifying small and densely distributed reproductive structures. Finally, YOLOv8 employs a decoupled detection head in which bounding box regression and classification are handled by separate branches, improving training stability and bounding-box localization accuracy.

3.3. Proposed Lightweight YOLOv8n

This section introduces a lightweight yet high-performance modification of YOLOv8n for unified multi-crop reproductive structure detection and counting across heterogeneous agri-

cultural datasets. The proposed architecture keeps the original backbone and detection head of YOLOv8n unchanged **?**, while modifications are applied only to the neck to improve multi-scale feature fusion. Target structures often share visual similarity with surrounding vegetation, appear under occlusion or overlap, and vary in size and density across crops and growth stages. However, the standard C2f blocks in YOLOv8 rely on conventional bottleneck structures that process spatial and channel information uniformly. To overcome this limitation, the neck is redesigned based on two main principles: (1) integrating attention mechanisms directly within cross-stage fusion to enable selective feature refinement during aggregation, and (2) improving feature interaction through structural re-parameterization rather than increasing network depth or width, thereby preserving computational efficiency. As illustrated in Figure 2, two coordinated modifications are introduced in the neck. First, the standard C2f blocks are replaced with C2f-SCA, an attention-enhanced cross-stage fusion module. Second, selected stride-2 convolution layers are replaced with RepDWConv, a re-parameterizable depthwise convolution operator.

3.4. Spatial–Channel Attention Convolution (SCA)

Figure 3 illustrates the structure of the proposed Spatial–Channel Attention Convolution (SCA) block. This module enhances discriminative feature representation while keeping the computational cost low. Given an input feature map $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W}$, the SCA block performs three sequential operations: (1) efficient feature mixing using a re-parameterizable depthwise–pointwise convolution, (2) parallel channel and spatial attention with adaptive gating, and (3) residual feature fusion with learnable scaling.

3.4.1. Re-parameterizable Depthwise Feature Mixing

The input feature tensor $\mathbf{X}$ is first processed using a re-parameterizable depthwise convolution (RepDWConv), followed by batch normalization, SiLU activation, and a 1×1 pointwise convolution:

$$\mathbf{Y} = \mathrm{PW}\left(\phi\left(\mathrm{BN}\left(\mathrm{RepDWConv}(\mathbf{X})\right)\right)\right), \tag{1}$$

where $\phi(\cdot)$ denotes the SiLU activation. During training, RepDWConv adopts a multi-branch structure consisting of a 3×3 depthwise convolution, a parallel 1×1 depthwise convolution branch, and an identity mapping:

$$\mathrm{RepDWConv}(\mathbf{X}) = \mathrm{BN}(\mathrm{DW}_{3\times3}(\mathbf{X})) + \mathrm{BN}(\mathrm{DW}_{1\times1}(\mathbf{X})) + \mathrm{BN}(\mathbf{X}), \tag{2}$$

The identity branch $\mathrm{BN}(\mathbf{X})$ is applied only when the convolution operates with stride $= 1$. For stride-2 downsampling configurations, the identity branch is omitted to maintain consistent spatial resolution across all branches before aggregation. At inference, these branches are structurally re-parameterized and fused into a single 3×3 depthwise convolution.

3.4.2. Channel–Spatial Attention

The mixed feature $\mathbf{Y}$ is refined using parallel channel and spatial attention mechanisms. Channel attention captures global contextual information by applying global average pooling to obtain a channel descriptor $\mathbf{z}$, which is passed through a lightweight transformation

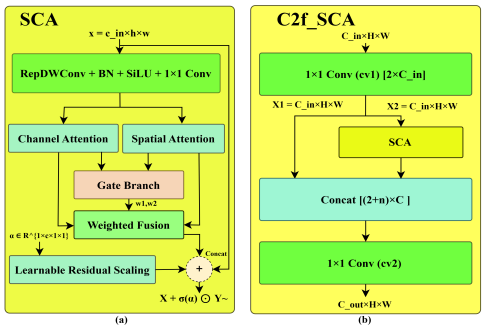

Figure 3: Architecture of the proposed (a) SCA, and (b) C2f-SCA.

consisting of two linear layers with SiLU activation and a sigmoid function to generate channel attention weights $\mathbf{a}_c$. The channel-refined feature is then computed as

$$\mathbf{Y}_{CA} = \mathbf{Y} \odot \mathbf{a}_c. \tag{3}$$

Spatial attention highlights informative regions using channel-wise average and maximum projections of $\mathbf{Y}$. These two maps are concatenated and processed by a 7×7 convolution followed by a sigmoid activation to produce spatial attention weights $\mathbf{a}_s$. The spatially refined feature is obtained as

$$\mathbf{Y}_{SA} = \mathbf{Y} \odot \mathbf{a}_s. \tag{4}$$

This dual attention captures both global channel dependencies and spatially salient regions.

3.4.3. Adaptive Gating and Residual Fusion

Instead of directly combining the two attention branches, SCA employs a dynamic gating mechanism to balance their contributions. A global descriptor is obtained by applying global average pooling to the mixed feature $\mathbf{Y}$ and passed through a lightweight transformation to produce gating logits. Let $\boldsymbol{\ell} \in \mathbb{R}^2$ denote the gating logits generated by the gating network. A softmax function then generates adaptive weights:

$$[w_1, w_2] = \mathrm{Softmax}(\boldsymbol{\ell}), \quad w_1 + w_2 = 1. \tag{5}$$

The fused feature is obtained as

$$\widetilde{\mathbf{Y}} = w_1 \mathbf{Y}_{CA} + w_2 \mathbf{Y}_{SA}. \tag{6}$$

Finally, residual fusion with learnable scaling produces the output feature:

$$\mathbf{O} = \mathbf{X} + \sigma(\boldsymbol{\alpha}) \odot \widetilde{\mathbf{Y}}, \tag{7}$$

Here, the pointwise convolution preserves the channel dimension so that $\mathbf{X}$, $\mathbf{Y}$, and $\widetilde{\mathbf{Y}} \in \mathbb{R}^{B \times C \times H \times W}$, enabling valid residual addition. The learnable scaling parameter $\boldsymbol{\alpha}$ is initialized to -2.0, resulting in a small initial attention contribution ($\sigma(-2) \approx 0.12$) that stabilizes early training. The sigmoid constraint gradually increases attention influence with minimal computational overhead.

3.5. Attention-Enhanced Cross-Stage Fusion (C2f-SCA)

To improve multi-scale feature aggregation in the neck, we introduce an attention-enhanced cross-stage fusion module termed C2f-SCA. The module integrates the proposed Spatial–Channel Attention Convolution (SCA) into the cross-stage partial fusion mechanism of YOLOv8n. The original C2f block improves gradient propagation through a split–transform–concatenate structure but relies on standard bottleneck blocks that do not explicitly emphasize salient spatial regions or informative feature channels. This limitation can reduce representation quality in dense agricultural scenes with occlusion and object overlap. C2f-SCA preserves the C2f topology while replacing bottleneck transforms with SCA modules, enabling attention-guided refinement during cross-stage fusion (Figure 3). Let the input feature map be

$$\mathbf{X} \in \mathbb{R}^{B\times C_{in}\times H\times W}. \tag{8}$$

where B, C_{in}, H, and W denote the batch size, channel dimension, height, and width. A 1×1 convolution first expands the channels:

$$\mathbf{Z} = \mathrm{Conv}_{1\times1}(\mathbf{X}) \in \mathbb{R}^{B\times 2C\times H\times W}. \tag{9}$$

The expanded feature map is split into two parts along the channel dimension, $\mathbf{X}_1$ and $\mathbf{X}_2$, where $\mathbf{X}_1$ acts as a bypass path, while the second branch is refined using n stacked SCA blocks:

$$[\mathbf{X}_1, \mathbf{X}_2] = \mathrm{Split}(\mathbf{Z}), \qquad \mathbf{X}_k = \mathrm{SCA}(\mathbf{X}_{k-1}), \quad k = 3, \ldots, n+2. \tag{10}$$

All intermediate features are concatenated along the channel dimension to form the fused representation $\mathbf{Z}_{cat}$. Finally, a 1×1 convolution produces the output feature:

$$\mathbf{Z}_{cat} = \mathrm{Concat}(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{n+2}), \qquad \mathbf{Y} = \mathrm{Conv}_{1\times1}(\mathbf{Z}_{cat}). \tag{11}$$

This design maintains compatibility with the original C2f structure while replacing bottleneck operations with SCA modules.

3.6. Experimental Setup and Evaluation Metrics

All experiments were implemented in PyTorch using the Ultralytics YOLOv8 framework on a single NVIDIA P100 GPU (16GB VRAM). Automatic Mixed Precision (AMP) was enabled to improve training efficiency and reduce GPU memory usage. The experiments follow the predefined training, validation, and test splits described in Section 3.1. All images were resized to 640×640 pixels. Data augmentation included mosaic augmentation (0.7 probability) and random translation (0.10 factor). The model was optimized using SGD (lr=0.005, momentum=0.937, weight decay=0.0005) with a cosine learning rate schedule and three warm-up epochs. Batch size was set to 8. Models were trained for 200 epochs on the Cauliflower and Tomato Flower datasets and for 300 epochs on the Maize Tassel dataset, with early stopping enabled (patience = 20).

Detection performance was evaluated using Precision, Recall, and mean Average Precision (mAP). Computational efficiency was evaluated using Frames per Second (FPS) and the number of model parameters. Counting performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination

Table 1: Detection performance comparison of different object detectors on the Cauliflower, Tomato Flower, and Maize Tassel datasets.

Model	Cauliflower					Tomato Flower					Maize Tassel					Params (M)
	P	R	F1	mAP@0.5	mAP@0.5:0.95	P	R	F1	mAP@0.5	mAP@0.5:0.95	P	R	F1	mAP@0.5	mAP@0.5:0.95	
Faster R-CNN	**0.953**	0.958	**0.955**	0.934	0.586	0.896	0.898	0.897	0.809	0.395	0.906	0.827	0.865	0.676	0.366	41.300
YOLOv5n	0.929	0.931	0.930	0.960	0.704	0.914	0.895	0.904	0.966	0.664	0.956	0.902	0.928	0.959	0.667	2.500
YOLOv6n	0.940	0.872	0.905	0.944	0.602	0.886	0.907	0.896	0.965	0.655	0.953	0.881	0.916	0.946	0.641	4.230
YOLOv7-tiny	0.947	0.962	0.954	0.977	0.669	0.934	0.897	0.910	0.961	0.666	**0.962**	0.901	0.931	0.954	0.596	6.000
RT-DETR-l	0.933	0.953	0.943	0.963	0.629	0.927	0.902	0.917	0.959	0.667	0.916	0.830	0.871	0.901	0.523	32.810
YOLOv9t	0.937	0.960	0.948	0.971	0.680	0.906	0.905	0.905	0.960	0.671	0.954	0.898	0.925	0.957	0.680	**2.010**
YOLOv10n	0.925	0.924	0.925	0.968	0.680	**0.937**	0.872	0.903	0.964	**0.668**	0.949	0.892	0.920	0.960	0.697	2.710
YOLOv8n	**0.953**	0.934	0.943	0.975	0.691	0.925	0.893	0.912	0.968	0.666	0.956	0.902	0.922	0.957	0.665	3.011
+ A + B (Ours)	0.928	**0.972**	0.954	**0.981**	**0.720**	0.934	**0.923**	**0.921**	**0.969**	0.677	**0.962**	**0.912**	**0.936**	**0.965**	**0.714**	2.659

(R^2). Object counts are estimated from the detection outputs by extracting confidence-based statistical features and predicting the count using an ExtraTrees regression model with linear calibration.

$$mAP = \frac{1}{C}\sum_{c=1}^{C} AP_c \tag{12}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i| \tag{13}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{14}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{15}$$

where y_i and $\hat{y}_i$ represent ground-truth and predicted counts; $\bar{y}$ is the mean ground-truth count; N is the number of samples; and C is the number of classes.

4. Results and Discussion

This section evaluates the proposed lightweight YOLOv8n with C2f-SCA and RepDWConv for multi-crop reproductive structure detection and counting across the Cauliflower, Tomato Flower, and Maize Tassel datasets.

4.1. Detection Performance Evaluation

Table 1 compares the detection performance of the proposed framework with representative detectors, including Faster R-CNN and several lightweight YOLO variants. On the Cauliflower dataset, the proposed lightweight YOLOv8n model (+A+B in Table 1) achieves the best performance with an mAP@0.5 of 0.981 and an mAP@0.5:0.95 of 0.720. These results improve upon the baseline YOLOv8n (0.975 and 0.691, respectively) while also achieving a higher F1-score of 0.954. Although the numerical improvement is modest, the consistent gain suggests improved stability when detecting small and densely clustered flowers, where preserving fine spatial information is critical. The Maize Tassel dataset presents additional challenges due to dense object distribution and frequent overlap. Under these

Table 2: Ablation study of RepDWConv (A) and C2f-SCA (B) on detection performance and inference speed across the Cauliflower, Tomato Flower, and Maize Tassel datasets. CBAM is included for comparison.

Model	C2f-SCA	RepDWConv	Cauliflower				Tomato Flower				Maize Tassel				Params (M)	GFLOPs
			mAP@0.5	mAP@0.5:0.95	F1	FPS	mAP@0.5	mAP@0.5:0.95	F1	FPS	mAP@0.5	mAP@0.5:0.95	F1	FPS		
YOLOv8n	×	×	0.975	0.691	0.943	**349**	0.968	0.666	0.912	**395**	0.957	0.665	0.922	**445**	3.011	8.1
+A	×	✓	0.975	0.699	0.948	326	0.967	**0.678**	0.917	346	0.958	0.675	0.928	272	2.996	8.5
+B	✓	×	0.976	0.701	**0.954**	332	0.968	**0.678**	0.920	372	0.959	0.679	0.927	375	2.819	8.3
+A+CBAM	×	✓	0.976	0.711	0.944	138	**0.970**	0.668	0.920	256	0.959	0.672	0.927	280	**2.649**	8.1
+A+B (ours)	✓	✓	**0.981**	**0.720**	**0.954**	224	0.969	0.677	**0.921**	278	**0.965**	**0.714**	**0.936**	264	2.659	**8.1**

Table 3: Ablation study of RepDWConv (A) and C2f-SCA (B) on counting performance across the Cauliflower, Tomato Flower, and Maize Tassel Datasets.

Model	C2f-SCA	RepDWConv	Cauliflower				Tomato Flower				Maize Tassel			
			MAE	RMSE	R^2	Acc(≤2)	MAE	RMSE	R^2	Acc(≤2)	MAE	RMSE	R^2	Acc(≤2)
YOLOv8n	×	×	0.374	0.749	0.961	0.986	0.225	0.543	0.951	0.995	1.205	1.895	0.994	0.883
+A	×	✓	0.281	0.715	0.959	0.986	0.245	0.533	0.955	**1.000**	1.111	1.754	0.995	0.879
+B	✓	×	0.281	0.715	0.957	0.986	**0.210**	0.519	0.955	**1.000**	1.098	1.706	0.995	**0.893**
+A+B	✓	✓	**0.258**	**0.709**	**0.965**	**0.988**	**0.210**	**0.489**	**0.962**	**1.000**	**1.068**	**1.679**	**0.996**	0.879

conditions, the proposed model achieves an mAP@0.5 of 0.965 and an F1-score of 0.936, compared with 0.957 and 0.922 for YOLOv8n. These improvements are achieved using only 2.659M parameters, which is significantly smaller than many larger detectors such as RT-DETR-L with over 32M parameters. Figure 4 compares the precision-recall performance of the baseline YOLOv8n and the proposed lightweight YOLOv8n across the three datasets. The proposed lightweight YOLOv8n achieves higher mAP@0.5 values, improving from 0.975 to 0.981 on Cauliflower, 0.968 to 0.969 on Tomato Flower, and 0.957 to 0.965 on Maize Tassel. The curves indicate more stable detection with higher precision across a wider recall range.

4.2. Ablation Study

Ablation experiments were conducted to evaluate the contributions of RepDWConv (A) and C2f-SCA (B) on the Cauliflower, Tomato Flower, and Maize Tassel datasets. Table 2 reports detection accuracy and inference speed for different model variants, while Table 3 summarizes the corresponding counting performance. The baseline YOLOv8n is used as the reference. On the Cauliflower dataset, mAP@0.5:0.95 increases from 0.691 to 0.699 and the F1-score improves from 0.943 to 0.948. On the Tomato Flower dataset, mAP@0.5:0.95 increases from 0.666 to 0.678. Similarly, on the Maize Tassel dataset, mAP@0.5:0.95 improves from 0.665 to 0.675 and the F1-score increases from 0.922 to 0.928. Replacing the standard C2f block with C2f-SCA (B) further improves detection accuracy by enabling Spatial–Channel Attention, allowing the model to focus on informative regions and suppress background interference. Combining both modules (A+B) yields the best detection performance across all datasets. More noticeable improvements are observed for mAP@0.5:0.95, which increases from 0.691 to 0.720 (4.2% gain) on Cauliflower, from 0.666 to 0.677 (1.65% gain) on Tomato Flower, and from 0.665 to 0.714 (7.4% gain) on Maize Tassel. The F1-

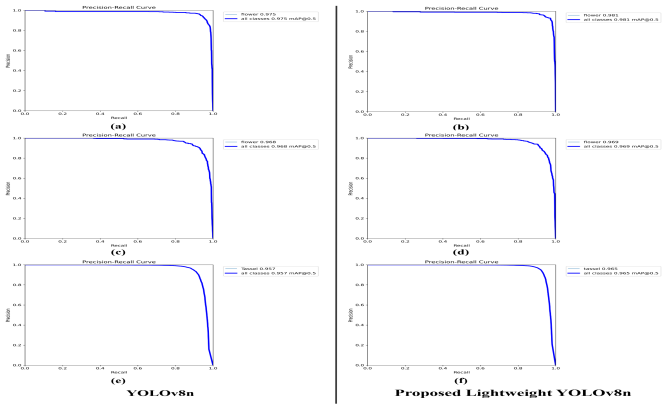

Figure 4: PR curves of YOLOv8n (baseline) and the proposed lightweight YOLOv8n on Cauliflower (a,b), Tomato Flower (c,d), and Maize Tassel (e,f) datasets.

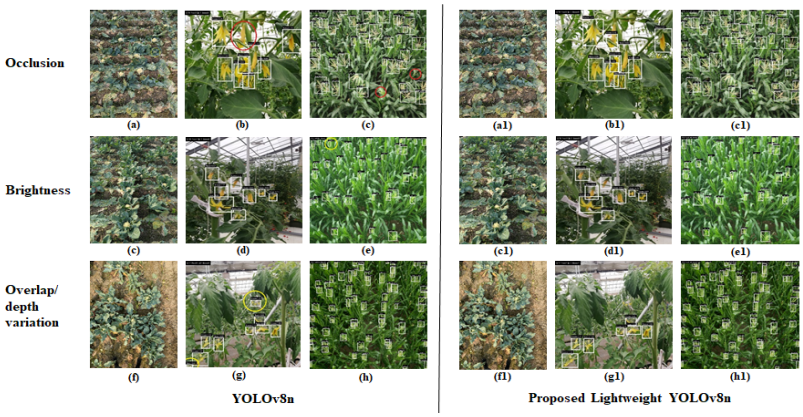

Figure 5: Qualitative comparison of YOLOv8n (left) and proposed model (right) on the same images from the Cauliflower, Tomato Flower, and Maize Tassel datasets under occlusion, brightness, and overlap conditions. Red circles denote false negatives and yellow circles denote false positives.

score also improves from 0.943 to 0.954 (1.2% gain), 0.912 to 0.921 (1.0% gain), and 0.922 to 0.936 (1.5% gain) across the three datasets. The proposed modules also improve counting performance. On the Cauliflower dataset, MAE decreases from 0.374 to 0.258 (31% reduction), accompanied by improvements in RMSE and R^2 (0.961 to 0.965). For the Tomato Flower dataset, RMSE decreases from 0.543 to 0.489 (9.9% reduction) while R^2 increases from 0.951 to 0.962, achieving perfect counting accuracy within an error tolerance of two objects. On the Maize Tassel dataset, the final model achieves the lowest MAE (1.068) and highest R^2 (0.996). Detection errors propagate to the counting stage, where false positives,

and false negatives lead to overestimation and underestimation, respectively. However, the proposed lightweight framework mitigates these effects, as evidenced by reduced MAE and real–time counting in complex agricultural environments. These results show that C2f-SCA and RepDWConv jointly enhance both detection and counting reliability while preserving a lightweight architecture suitable for real-time agricultural applications.

4.3. Qualitative Evaluation

Figure 5 shows qualitative comparisons between YOLOv8n and the proposed lightweight YOLOv8n on the same images from the Cauliflower, Tomato Flower, and Maize Tassel datasets. Baseline errors are marked by red circles (false negatives) and yellow circles (false positives), while the proposed model detects more reproductive structures in complex scenes.

5. Conclusion

This study proposes a lightweight YOLOv8n-based framework for unified multi-crop reproductive structure detection and counting. The method enhances feature representation and multi-scale fusion by integrating a Spatial–Channel Attention Convolution module within cross–stage fusion (C2f-SCA) and a re-parameterizable depthwise convolution (RepDWConv). Experiments conducted on three heterogeneous datasets: Cauliflower, Tomato Flower, and Maize Tassel demonstrate the effectiveness of the proposed design. The model achieves mAP@0.5 of 0.981 and mAP@0.5:0.95 of 0.720 on the Cauliflower dataset, while requiring only 2.659M parameters and reducing MAE by up to 31%. Overall, the proposed framework provides an efficient and scalable solution for automated crop monitoring, yield estimation, and scalable precision agriculture through multi-crop reproductive structure detection and counting in resource-constrained agricultural environments. A limitation of this study is evaluation on only three crop datasets; future work will validate the framework on larger multi-crop datasets and more diverse field conditions.

Data Availability Statement

The cauliflower dataset introduced in this study is available at: https://zenodo.org/records/20011067. The tomato flower and maize tassel datasets used in this study are publicly available from their original sources as cited in this paper.

Acknowledgments

The authors sincerely acknowledge the local farmers of Bogura and Dhaka, Bangladesh, for granting access to their fields and supporting data collection.