CenterFormer: Coupling CenterNet and Vision Transformer for Accurate Wheat Head Detection

Ekei Harimoto and Xian-Hua Han *

Graduate School of Artificial Intelligence and Science, Rikkyo University, Tokyo, Japan
Email: 24vr032p@rikkyo.ac.jp (E.H.); hanxhua@rikkyo.ac.jp (X.-H.H.)
*Corresponding author

***Abstract*—Wheat is a staple crop cultivated widely across the world, making effective management of wheat fields a critical task. A key component of this management is accurately identifying and counting wheat heads, which provides essential data for assessing growth conditions, estimating crop yields and optimizing agricultural. This study introduces a novel approach for automatic wheat head detection by treating the wheat head as a single point to avoid ambiguous annotation of dense objects while leveraging the long-range dependency modeling capabilities of Transformer architecture to learn multi-scale features for head prediction, dubbed as CenterFormer. Specifically, we employ a hierarchical Transformer architecture with self-attention exploitation in both spatial and channel domains as the backbone to extract multi-scale features in the hierarchical stages. To maintain the linear complexity of the Transformer block, we implement window-based self-attention in spatial domain and group-wised self-attention in channel direction. In addition, to leverage the multi-scale features with both detailed spatial information and abstracted semantic contexts, we design a simple yet effective fusion block to integrate these features for enhanced wheat prediction. The prediction block aims to estimate a heat map, denoting the probabilities if the points are located at the centers of the wheat heads, and regresses other object properties such as size and sub-pixel deviations for each center location. Extensive experiments on the Global Wheat Head Detection (GWHD) dataset have demonstrated that our proposed method achieves substantial performance improvements compared with the state-of-the-art object detection models.**

***Keywords*—wheat head detection, transformer, self-attention, multi-scale feature fusion, hierarchical architecture, center point, CenterNet**

I. Introduction

Wheat serves as a fundamental dietary crop for approximately 30% of the global population, highlighting its critical role in global food security [1]. As the world's population continues to grow, the demand for increased crop production becomes more pressing. Enhancing wheat productivity is expected to have a substantial impact on the global food supply, necessitating innovative agricultural optimization strategies to meet future needs [2, 3]. Among the optimization way, effective management of wheat fields has emerged as a key focus for boosting production yields. One pivotal and widely used strategy in assessing wheat field situations is the wheat head count, which provides valuable insights into crop growth dynamics. The density of wheat heads during the growth stage is particularly significant, as it serves as a direct predictor of potential yield [4]. However, accurately measuring wheat head density presents numerous challenges. The variability in wheat head orientation, often influenced by environmental factors such as wind, the relatively small size and their overlapping clusters, make manual counting both labor-intensive and prone to error. To address these challenges, the adoption of automated detection methods is essential. Leveraging advances in machine learning and image processing technologies, these automatic methods have potential of facilitating accurate and efficient wheat head detection, enabling scalable and timely management of crop fields, and thus attracted substantial research attention for supporting efforts to optimize agricultural productivity [5–7].

In recent years, Deep Convolutional Neural Networks (DCNNs) [5–7] have emerged as impressive tools in computer vision, achieving significant advancements across a wide range of tasks including object detection. Many deep models such as the region-based CNN (R-CNN) series [8–11], single shot multibox detector (SSD) [12], and you look only once (YOLO) models [13–17] have been extensively exploited for general object detection, and have demonstrated promising performance on publicly available image datasets [18, 19], solidifying their position as state-of-the-art solutions in the field. These methods require to generate large number of bounding boxes as object candidates based on predefined anchor points, and thus are usually categorized as anchor-based methods. Although such approaches have been employed in wheat head detection tasks [20–23], their design is inherently optimized for detecting generic objects, typically characterized by larger sizes and more distinct spatial locations. Consequently, directly applying these anchor-based methods to wheat head detection often results in significant performance degradation, particularly for small-sized objects like wheat heads. To address this challenge, Feature Pyramid Networks (FPN) [24] have

Manuscript received January 17, 2025; revised April 7, 2025; accepted May 26, 2025; published September 17, 2025.

been integrated into detection models to improve performance across objects of varying sizes. Despite these enhancements, most existing methods aggregate multi-scale features with heavily down-sampled spatial resolutions, limiting their ability to preserve the intricate details needed to detect small objects, such as wheat heads. This limitation is exacerbated by the dense and overlapping nature of wheat heads, which complicates the generation of precise annotations for the large number of bounding boxes required. As a result, achieving accurate and efficient wheat head detection remains a substantial challenge, necessitating novel strategies to account for these unique characteristics.

In contrast, several anchor-free object detection models that conceptualize objects as single points have been proposed [25–27], demonstrating remarkable performance. For instance, Zhou *et al.* introduced an approach that focuses exclusively on the object center as the positive candidate, referred to CenterNet [26]. They represented the object center point as a heatmap to effectively detect objects in highly crowded scenarios with significant overlap. However, its performance is constrained when dealing with small objects, due to its reliance on significantly downsampled feature maps during detection. This limitation arises because the reduced spatial resolution impairs the model's ability to capture fine-grained details necessary for accurately detecting small-sized objects.

Besides, conventional detection models typically utilize CNN architectures as their backbone. While these architectures are adept at capturing local features, they often struggle to effectively model long-range dependencies. In the context of wheat head detection, the field environment is characterized by densely distributed plants and complex structural arrangements. Leveraging the global contextual correlations among multiple wheat heads can facilitate the extraction of more semantic and intrinsic representations, thereby enhancing detection performance. Recently, Transformer architectures, which is firstly proposed in Natural Language Processing (NLP) [28], has been introduced as a novel paradigm to capture global dependencies, prompting researchers to adapt these models for computer vision applications [29–34]. Unlike the feature abstraction of local receptive fields in the CNN architectures, Transformers inherently produce global receptive fields, making them more effective for object detection in crowed and complex environments. The dominated component in the Transformer block leverages the attention mechanism to capture global dependencies, potentially achieving more semantic and intrinsic representations of the focused objects for accurate localization.

To this end, this study proposes a novel method for automatic wheat head detection, referred to as CenterFormer, which represents each wheat head as a single point to mitigate the challenges of ambiguous annotations in densely populated scenarios. The approach leverages the Transformer architecture's capacity for modeling long-range dependencies to learn multi-scale features critical for accurate head prediction. Specifically, a hierarchical Transformer architecture is employed as the backbone, utilizing self-attention mechanisms in both spatial and channel dimensions to extract multi-scale features across hierarchical stages. To ensure the computational efficiency of the Transformer block, window-based self-attention is applied in the spatial domain, while group-wise self-attention is implemented in the channel domain, maintaining linear complexity. Moreover, we explore a cross-scale attention mechanism to refine the low-level features in the shallow stages according to the high-level semantic feature of the final stage. Finally, a simple yet effective fusion block is introduced to integrate multi-scale refined features, combining detailed spatial information with abstracted semantic contexts to enhance predictive performance. The prediction module generates a heat map that represents the likelihood of points corresponding to the centers of wheat heads. Additionally, it also estimates the object size of each center location, providing a comprehensive representation for wheat head detection. Comprehensive experiments conducted on the Global Wheat Head Detection (GWHD) dataset validate the efficacy of the proposed method, showcasing significant performance enhancements over state-of-the-art object detection models.

In summary, the key contributions of this study can be outlined as follows:

1) We introduce CenterFormer, a novel framework for accurate wheat head detection that combines the Transformer architecture's robust capability for modeling long-range dependencies with the CenterNet framework's effectiveness in predicting distinct objects in densely populated scenarios

2) We utilize a dual-attention Vision Transformer to exploit correlations in both spatial and channel domains, dubbed as Spatial and Channel Attention (SCA) based Transformer, and preserve multi-scale features across hierarchical stages, facilitating the detection of small-sized wheat heads.

3) A cross-scale attention mechanism is implemented to refine low-level features, incorporating detailed spatial information from shallow stages with high-level semantic features from deeper stages. This mechanism enables the extraction of multi-scale, discriminative, and intrinsic representations of wheat heads.

4) We design a simple yet effective fusion block to integrate refined multi-scale features, effectively combining detailed spatial information with abstracted semantic contexts, thereby improving predictive performance.

II. Literature Review

This section firstly presents a brief survey of the existing detection models for the wheat heads, and then introduces the closely related techniques to our proposed model, including the anchor-free based CenterNet and vision Transformer architecture.

A. Existing Deep Models for Wheat Head Detection

Various deep learning models for generic object detection, such as the R-CNN series [8–11], SSD [12], and YOLO models [13–17], have been proposed and achieved remarkable success across diverse applications. Recent studies for wheat head detection have focused on adapting these advanced models to the specific domain tasks. For instance, Hasan *et al.* [35] employed four variations of the R-CNN model to detect wheat spikes and evaluate yields across different wheat varieties. Madec *et al.* [36] explored two complementary approaches, employing the Faster R-CNN network and the TasselNet local count regression network, to estimate wheat spike density using high-resolution RGB images. Gong *et al.* [20] advanced the YOLOv4 framework by integrating a Dual Spatial Pyramid Pooling (SPP) module, developing a highly efficient real-time detection system for wheat spikes. Similarly, Yang *et al.* [37] incorporated the Convolutional Block Attention Module (CBAM) [38, 39] into the CNN backbone within the YOLOv4 framework, improving detection performance through enhanced attention mechanisms. Further advancements include the work of Sun *et al.* [40], who introduced WHCnet, an enhanced wheat head counting network leveraging an improved feature pyramid network (AugFPN) to address challenges related to low detection accuracy. Ye *et al.* [41] developed WheatLFANet, a lightweight, real-time neural network optimized for efficient wheat head detection and counting, particularly on low-resource devices. Yan *et al.* [42] applied the GradCAM interpretability technique to refine detection layer scales in wheat spike detection networks. Additionally, Zhao *et al.* [43] proposed WheatNet, a model designed to detect wheat spikes throughout developmental stages, from the filling to maturity phases. These studies demonstrate the potential of adapting generic object detection models, and possible performance improvement with the incorporation of the domain-specific modifications according to the characteristic of the wheat head detection scenario. However, all the above methods employed the CNN architectures, which possess a strong capability for capturing local features while exhibit limitations in capturing global context. These limitations become particularly evident in the crowed and complex scenarios associated with wheat spike detection. In generic object detection application, detection Transformer (DETR) [44] has been proposed to achieve the powerful modeling capability of long-range dependencies, and manifested great performance improvement. Recently, Yang *et al.* [45] proposed to combine the DETR and a lightweight feature pyramid for wheat spike detection in complex background.

B. CenterNet

CenterNet [26] is a cutting-edge object detection framework that identifies objects by predicting their geometric centers, simplifying the detection process while maintaining high accuracy. Unlike conventional deep models [8–12, 14–16] that rely on anchor boxes or region proposals, CenterNet directly predicts heatmaps representing object centers, along with additional regression tasks for object size and offsets. This anchor-free approach eliminates the need for complex post-processing steps and reduces computational overhead. Therefore, the simplified pipeline of the CenterNet framework consolidates object detection tasks into a unified framework, reducing the need for separate region proposals or Non-Maximum Suppression (NMS). By predicting object centers and other properties (e.g., dimensions and offsets) in a single stage, CenterNet achieves competitive performance while being computationally efficient. Benefiting from the anchor-free design of the CenterNet, it particularly well-suited for tasks involving densely packed objects, such as wheat head detection. By focusing on the center of each wheat head, this framework demonstrates the following advancements. 1) It reduces the ambiguity of overlapping objects; 2) It enhances detection performance in crowded and occluded scenarios; 3) It simplifies the annotation process, as only object centers need to be labeled. Taking the previously discussed attributes of the CenterNet framework into account, this study utilizes the CenterNet pipeline to facilitate wheat head detection.

C. Vision Transformer

The Transformer architecture, introduced by Vaswani *et al.* [28], has become a cornerstone in Natural Language Processing (NLP) tasks. More recently, Dosovitskiy *et al.* [29] extended this framework to computer vision with the introduction of the Vision Transformer (ViT). Unlike traditional CNNs, the Transformer possesses a significantly larger receptive field and performs feature aggregation based on relationships directly learned from pairwise feature interactions, and has become a powerful and competitive module for feature learning in computer vision applications. Numerous studies have been done to develop more advanced Transformer-based networks for applications such as image, classification [29–34], object detection [44, 46], and semantic segmentation [47, 48]. To further enhance the representation capability, various efforts have been done for vision tasks, and made significant progress. For instance, Swin Transformer [30, 31] incorporate a hierarchical architecture with shifted windows for attention, improving scalability and computational efficiency while maintaining global context awareness. Touvron *et al.* [32] proposed to fucus on data-efficient training to achieve strong performance with smaller datasets while Ding *et al.* [49] explored a Dual attention Vision Transformer (DaViT) by combining the self-attention mechanisms in both spatial and channel domains for image classification. In object detection scenario, DETR (Detection Transformer) [44] employs a Transformer encoder-decoder framework to directly predict objects from learned query embeddings, providing a novel approach to object detection. Building on this, Deformable DETR [44] introduces a deformable attention mechanism to enhance training efficiency and convergence. Recent studies [50–52] have highlighted that incorporating guidance mechanisms, such as anchor boxes, can substantially enhance the convergence speed and stability of DETR-based models. Despite the notable advancements obtained by these models, they continue to

face significant challenges, including high computational costs, particularly when processing large input images. Additionally, their require long training times and large datasets to achieve convergence, primarily due to the inherent difficulties in learning object representations and the complexities associated with the bipartite matching process. Recently, several studies have explored Transformer-based approaches for wheat head detection [45, 53 54]. For instance, Yang *et al.* [45] adopted a DETR-style framework by integrating a CNN backbone with a Transformer encoder-decoder architecture, dubbed as WH-DETR. This design eliminates the need for hand-crafted components such as anchor generation and the Non-Maximum Suppression (NMS) step. Basically, WH-DETR inherits the limitations of the original DETR, such as slow convergence and difficulty in detecting densely packed objects. Zhou *et al.* [53] exploited a WheatFormer by employing a multi-window Swin Transformer as the feature extraction backbone. While the WheatFormer is effective in modeling local spatial dependencies through window-based attention, it is limited in capturing global context, which is essential for robust wheat head detection across diverse and complex field conditions. Additionally, Suma *et al.* [54] proposed a two-stage architecture, termed CETR, wherein a Vision Transformer (ViT) was employed as the second stage to enhance the feature representations extracted by a CenterNet-based backbone in the first stage. The CETR demonstrated impressive improvements in wheat head detection performance compared to CNN backbones. However, CETR's architecture involved a naïve concatenation of CenterNet and ViT modules, resulting in substantial computational overhead.

In this study, we aim to harness the superior capability of Transformers in modeling long-range dependencies by leveraging the recently developed Dual Attention Vision Transformer (DaViT) [49], which is designed with Spatial window and Channel group Attention (SCA) to achieve computational efficiency without compromising representational power. Specifically, we propose to integrate DaViT as the backbone within the CenterNet framework, thereby unifying the SCA mechanisms of DaViT with a streamlined detection pipeline, which offers the following advantages over standard ViT-based and DETR-like architectures. 1) Linear complexity is achieved via a hybrid of window-based and group attention mechanisms, significantly reducing computational cost compared to global self-attention. 2) The SCA module enables effective multi-scale feature extraction and spatially-aware context aggregation, facilitating both local and global dependency modeling. 3) These design choices result in faster convergence and more accurate localization, particularly in dense and cluttered scenes typical of in-field wheat head images.

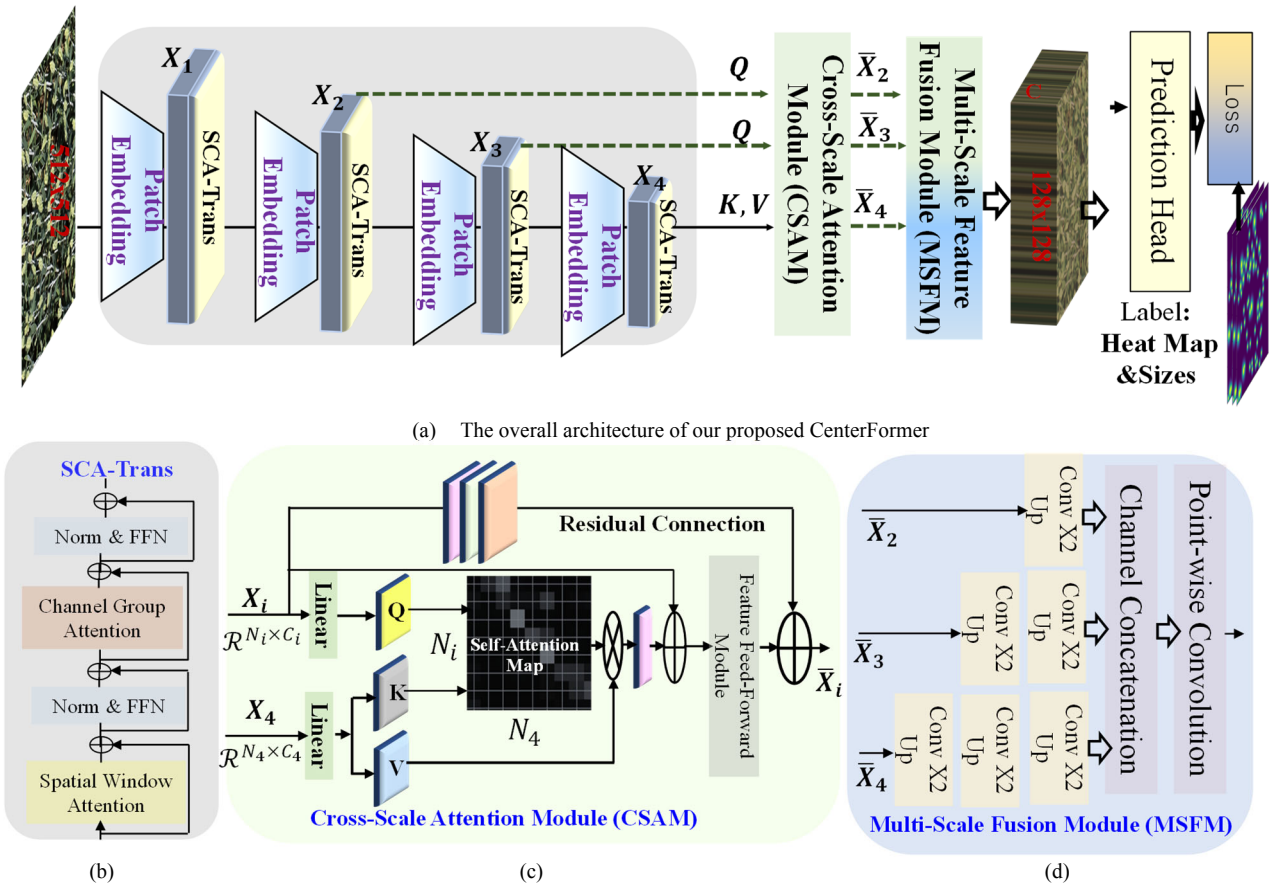
III. Proposed Method

This section firstly presents the overview of proposed CenterFormer for the wheat head detection, and then introduces the contributed components in our detection pipeline.

A. Overview

This study aims to exploit a novel CenterFormer by leveraging the complementary strengths of two advanced methodologies: the Transformer's powerful capacity for capturing long-range dependencies and modeling complex relationships in data, and the simplified, efficient detection framework provided by CenterNet. This combination is specifically designed to address the complex challenges inherent in detecting wheat heads within intricate agricultural environments. These challenges include handling occlusions, where spikes are partially obscured, and managing overlaps, where multiple wheat heads are closely clustered, making accurate detection particularly demanding. The overall structure of our proposed CenterFormer is manifested in Fig. 1(a).

Specifically, our approach employs a dual-attention Transformer block as the core component of the representation learning backbone, which operates across both spatial and channel domains, dubbed as SCA-Trans block. This design enables the model to capture rich contextual dependencies and multi-scale features, ensuring robust feature representation. Similar to the encoder structures commonly adopted in state-of-the-art image classification models [30, 31, 34, 55], the backbone is organized in a hierarchical layout. This layout begins with an initial patch embedding layer, which partitions the input image into fixed-size patches and embeds them into a high-dimensional feature space. Subsequently, the backbone consists of four progressive stages, each designed to extract representative features at varying spatial resolutions. Given an image $I \in \mathcal{R}^{W\times H\times 3}$, the backbone encoder hierarchically extracts representative features across four stages, each corresponding to progressively coarser spatial resolutions and richer semantic abstractions. Specifically, these features are represented as $X_1 \in \mathcal{R}^{\frac{W}{4}\times\frac{H}{4}\times C_1}$, $X_2 \in \mathcal{R}^{\frac{W}{8}\times\frac{H}{8}\times C_2}$, $X_3 \in \mathcal{R}^{\frac{W}{16}\times\frac{H}{16}\times C_3}$, and $X_4 \in \mathcal{R}^{\frac{W}{32}\times\frac{H}{32}\times C_4}$, transitioning from fine-grained spatial details in the early stages to global semantic context in the later stages. This hierarchical feature extraction enables the model to effectively represent information at varying scales. Given that wheat heads are typically small-sized and densely distributed, both fine-grained spatial details and high-level semantic contexts play a critical role in their accurate detection and localization. Fine-grained details help identify subtle features of individual wheat heads, while semantic context aids in distinguishing them from background and overlapping regions. To address this, we integrate multi-scale feature representations across the encoder's stages. This approach facilitates the generation of a robust and discriminative predictive feature map, effectively capturing the intricate balance of local detail and global context necessary for precise wheat head localization.



(a) The overall architecture of our proposed CenterFormer

(b) (c) (d)

Fig. 1. Conceptual scheme of the proposed CenterFormer. (a) The overall architecture of the CenterFormer; (b) The SCA-Trans block; (c) The cross-scale attention module (CSAM); (d) The multi-scale fusion module (MSFM).

To enhance the representation learning process, we design and investigate a Cross-Scale Attention Module (CSAM) that integrates information from different stages of the backbone. Specifically, the CSAM refines the low-level, fine-grained features extracted from earlier stages by incorporating the global semantic context captured in the final stage. This process can be mathematically expressed as $\bar{X}_i = f_{CSAM}(X_i, X_4)$, where $\bar{X}_i$ represents the refined feature map for stage represents the refined feature map for stage i, and f_{CSAM} denotes the cross-scale attention operation. By aligning and combining fine spatial details with high-level semantic information, this module enables the resulting representations to encode both precise spatial structures and enriched semantic content. Following the refinement of features at individual scales, the representations from multiple stages $(\bar{X}_2, \bar{X}_3, \bar{X}_4)$ are aggregated using a Multi-Scale Fusion Module (MSFM). This module performs an efficient yet effective fusion of the refined features to produce a unified predictive feature map $\bar{X} = f_{MSFM}([\bar{X}_2, \bar{X}_3, \bar{X}_4])$. The MSFM is designed to preserve complementary information across scales, allowing the predictive feature map to simultaneously capture detailed local features and broad contextual cues.

In our framework, the predictive feature map is produced with horizontal and vertical dimensions of $\frac{W}{4}$ and $\frac{H}{4}$, respectively. Building upon this feature map, we employ two separate prediction heads, inspired by the CenterNet pipeline, to directly estimate the outputs. The first head generates a center heatmap $H \in \mathcal{R}^{\frac{W}{4}\times\frac{H}{4}\times 1}$, which represents the likelihood of object centers at each spatial location. The second head perform box regression $B \in \mathcal{R}^{\frac{W}{4}\times\frac{H}{4}\times 2}$, which encodes the horizontal and vertical sizes of the bounding boxes for detected objects. Since the wheat head detection scenario involves a single object class, the center heatmap H is designed with a single channel while the box regression B focuses solely on the spatial dimensions of the objects, minimizing computational complexity. During training, the ground truth for the center heatmap is derived from a Gaussian function centered on the annotated bounding box centers, effectively guiding the model to learn precise localization. This design ensures an efficient and effective prediction pipeline tailored to the specific requirements of wheat head detection in agricultural settings. The overall architecture of our proposed CenterFormer is manifested in Fig. 1.

In the subsequent sections, we detail the architectural components of our proposed model, including the hierarchical backbone constructed with CSA-Transformer blocks, which serve as the foundational building units for representation learning. Additionally, we present the Cross-Scale Attention Module (CSAM), designed to refine features across different scales by leveraging both fine-grained and semantic information. The Multi-Scale Fusion Module (MSFM) is then introduced as an effective

mechanism for integrating features from multiple scales to produce a unified representation.

B. The Hierarchical SCA-Transformer Backbone

The SCA-Transformer backbone is organized into a hierarchical structure comprising four distinct stages, each designed to progressively extract features of increasing semantic complexity while preserving essential spatial and channel-wise information. At the beginning of each stage, a patch embedding layer is introduced to transform the input features into a structured representation suitable for subsequent processing. This embedding operation reduces the spatial resolution while increasing the feature dimensionality, enabling efficient computation and better feature abstraction.

Within each stage, multiple Spatial and Channel Attention (SCA) blocks are stacked to enhance feature learning. These blocks operate at a fixed resolution and feature dimensionality throughout the stage, ensuring that the spatial details and channel dependencies are effectively captured and retained. Given the input image $I \in \mathcal{R}^{W\times H\times 3}$, the overall backbone obtains the hierarchical features, with progressively decreasing spatial resolutions and increasing channel dimensions. The feature resolutions of 4 stages are $\frac{W}{4}\times\frac{H}{4}$, $\frac{W}{8}\times\frac{H}{8}$, $\frac{W}{16}\times\frac{H}{16}$, and $\frac{W}{32}\times\frac{H}{32}$ while the corresponding channel dimensions are C, $2C$, $4C$, and $8C$, respectively. The patch embedding layers, responsible for initializing the transformation at the start of each stage, are implemented using stride convolution operations. These layers reduce the spatial resolution while increasing the channel dimensionality to encode richer feature representations. The convolutional kernels and stride values used in the four patch embedding layers are $\{7, 2, 2, 2\}$ and $\{4, 2, 2, 2\}$, respectively, ensuring effective downsampling and feature encoding. In our experiments, the base channel dimension C is set to 96. The SCA-Transformer blocks, which serve as the primary computational units within each stage, are configured with a stage-specific number of blocks: $\{1, 1, 9, 1\}$.

In detail, the core attention mechanism within the CSA-Transformer block is realized through the integration of two specialized components: a spatial window attention block and a channel group attention block. Given the input feature representation, we first implement a spatial Transformer that leverages spatial window attention to capture spatial dependencies and relationships within the feature map. Subsequently, we proceed with a channel Transformer, which utilizes channel group attention to model inter-channel dependencies. Together, these two stages form the core of the SCA-Trans block. This block is designed to synergistically integrate spatial and channel-wise contextual information, enhancing the expressive power and representational capability of the input features. Fig. 1(b) illustrates the architecture of the CSA-Trans block.

Spatial window attention: Given a feature map F, we perform a window partitioning operation to divide F into M local patches, with the feature representation of the $m-th$ patch denoted as F_m. Let us assume that $F_m \in \mathcal{R}^{P\times C}$, where P represents the number of tokens within each patch and C denotes the total number of feature channels. To model the internal dependencies within each patch, we employ the standard Multi-Head Self-Attention (MHSA) mechanism. This mechanism is mathematically defined as follows:

$$\mathcal{A}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Concat}\left(\text{Head}_1,\ldots,\text{Head}_H\right) \tag{1}$$

where each attention head Head_h is computed as:

$$\begin{aligned}\text{Head}_h &= \text{Atention}(\mathbf{Q_h},\mathbf{K}_h,\mathbf{V}_h)\\ &= \text{Softmax}\left[\frac{\mathbf{Q_h}(\mathbf{K}_h)^T}{\sqrt{C_h}}\right]\mathbf{V}_h,\end{aligned} \tag{2}$$

where $\boldsymbol{Q_h}$, $\boldsymbol{K_h}$, and $\boldsymbol{V_h}$ are query, key, and value matrices of the $h-th$ head, respectively. These matrices are defined as:

$$\mathbf{Q_h} = \mathbf{F_m^h}\,\mathbf{W_h^Q},\ \mathbf{K_h} = \mathbf{F_m^h}\,\mathbf{W_h^K},\ \text{and}\ \mathbf{V_h} = \mathbf{F_m^h}\,\mathbf{W_h^V} \tag{3}$$

where $\boldsymbol{W_h^Q}$, $\boldsymbol{W_h^K}$, and $\boldsymbol{W_h^V}$ are learnable weight matrices, and $\boldsymbol{F_m^h}$ refers to the input feature of the $h-th$ attention head derived from $\boldsymbol{F}_m$. Each head processes a feature of dimensionality $\mathcal{R}^{P\times C_h}$, where C_h is the number of channels per head. It is important to note that the total number of channels, C, is related to C_h by the equation $C = H\times C_h$, where H is the total number of attention heads. After applying the MHSA operation independently to each patch, the resulting output features from all patches are aggregated to form the final feature representation.

The spatial window attention block focuses on capturing local dependencies by partitioning the feature map into non-overlapping spatial windows, allowing the model to compute self-attention within each window efficiently. This approach ensures that spatial relationships are preserved while maintaining computational feasibility for high-resolution feature maps.

Channel group attention: In computer vision, self-attention mechanisms are commonly employed to capture relationships between image tokens, where tokens are typically defined as individual pixels or small patches. These methods primarily gather information across spatial dimensions to model spatial dependencies effectively. Expanding upon this, we incorporate an attention mechanism to model complementary dependencies in the channel domain. This approach treats each channel as a distinct token, reshaping the feature map of each channel into a single vector. These channel tokens are then processed to interact with global information along the channel dimension. Importantly, this interaction is achieved with linear complexity concerning spatial dimensions, making it computationally efficient.

To further optimize computational efficiency, we group the channels into multiple subsets and perform self-attention within each group independently. Formally, let G denote the number of groups and C_g the number of channels per group, such that the total number of channels is given by $= G\times C_g$. By structuring the attention in this way, the channel group attention mechanism remains global, enabling image-level token interactions across a

specific group of channels. The mathematical formulation of the channel group attention mechanism is as follows:

$$\mathcal{A}_{CH}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \left\{\mathcal{A}_g\left(\mathbf{Q}_g,\mathbf{K}_g,\mathbf{V}_g\right)\right\}_{g=1}^{G} \quad (4)$$

where the attention operation for the g-th group, $\mathcal{A}_g$, is defined as:

$$\mathcal{A}_g\left(\mathbf{Q}_g,\mathbf{K}_g,\mathbf{V}_g\right) = softmax\left[\frac{\mathbf{Q_g}(\mathbf{K}_g)^T}{\sqrt{C_g}}\right]\mathbf{V}_g \quad (5)$$

where $\boldsymbol{Q_g}$, $\boldsymbol{K_g}$, and $\boldsymbol{V_g}$ represent the channel-wised image-level query, key, and value matrices of the $g-th$ group, respectively. These matrices are derived from the feature representations of the channels in the $g-th$ group. This formulation ensures that the attention mechanism operates efficiently and effectively captures global channel interactions, while computational complexity is reduced by confining attention computations to individual channel groups.

C. Cross-Scale Attrntion Module: CSAM

As previously introduced, the CSA-Transformer backbone generates multi-scale feature maps through its four hierarchical stages. Each stage corresponds to progressively coarser spatial resolutions and captures increasingly rich semantic abstractions. Specifically, we use the feature map from stage 4 ($\boldsymbol{X}_4$) for refining the features from all preceding stages because it captures high-level, semantically rich information extracted from deeper layers. By interacting with intermediate features ($\boldsymbol{X}_i$), $\boldsymbol{X}_4$ can provide global contextual cues that help to enhance important structures and suppress noise in earlier-stage features. This cross-scale interaction improves the overall discriminative capability of the network. In addition, since $\boldsymbol{X}_4$ has a much smaller spatial resolution compared to other $\boldsymbol{X}_i$, using it as the base Key and Value in CSFM significantly reduces the computational cost. Formally, let $\boldsymbol{X}_i \in \mathcal{R}^{N_i \times C_i}$ represent the feature map from the $i-th$ stage, where N_i denotes the number of spatial tokens and C_i represents the channel dimension. Similarly, let $\boldsymbol{X}_4 \in \mathcal{R}^{N_4 \times C_4}$ denote the feature map of the fourth stage. To facilitate the refinement of $\boldsymbol{X}_i$, we extract the query, key and value representations from $\boldsymbol{X}_i$ and $\boldsymbol{X}_4$. These are defined as follows:

$$\mathbf{Q_i} = \mathbf{X}_i\,\boldsymbol{W^Q}, \mathbf{K} = \mathbf{X}_4\,\boldsymbol{W^K},\ \mathbf{V} = \mathbf{X}_4\,\boldsymbol{W} \quad (6)$$

where $\boldsymbol{W^Q}$, $\boldsymbol{W^K}$, and $\boldsymbol{W^V}$ are learnable weight matrices used to project the input features into the query, key, and value spaces, respectively. To refine the feature $\boldsymbol{X}_i$, we compute Cross-Scale Attention (CSA) using the query $\boldsymbol{Q}$ from stage i and the key-value pair $(\mathbf{K}, \mathbf{V})$ from stage 4. The cross-scale attention is mathematically expressed as:

$$\mathcal{A}_{CSA}(\mathbf{Q}_i,\mathbf{K},\mathbf{V}) = \text{softmax}\left[\frac{\mathbf{Q}_i(\mathbf{K})^T}{\sqrt{C}}\right]\mathbf{V} \quad (7)$$

where softmax applies a normalization operation to the attention scores, and the resulting weighted sum of $\boldsymbol{V}$ provides a refined representation of the feature of the $i-th$ stage.

Following the computation of cross-scale attention, we further process the refined feature using a Feed-Forward (FF) transformation, implemented via a simple Multi-Layer Perceptron (MLP) subnetwork. The final refined feature $\overline{\boldsymbol{X}}_i$ is obtained as:

$$\overline{\boldsymbol{X}}_i = f_{CSAM}(\boldsymbol{X}_i,\boldsymbol{X}_4) = FF\left(LM(\mathcal{A}_{CSA} + \boldsymbol{Q}_i)\right) + \boldsymbol{Q}_i \quad (8)$$

where LM denotes a layer normalization operation, and the residual connection ($+\boldsymbol{Q}_i$) ensures stability and preserves the original feature information. The detailed architecture of the CSAM is demonstrated in Fig. 1(c).

This refinement process enables effective integration of the rich semantic information from the fourth stage with the multi-scale feature maps from preceding stages, enhancing the overall representational capacity of the network.

D. Multi-scale Feature Fusion Module: MSFM

To enhance the refined multi-scale feature representations, we introduce a simple yet effective fusion module called the Multi-Scale Feature Fusion Module (MSFM). This module is designed to integrate multi-scale features into a unified representation that combines detailed spatial information with rich semantic context. The detailed architecture of the MSFM is demonstrated in Fig. 1(d).

Particularly, the MSFM operates on the hierarchical features $\overline{\boldsymbol{X}}_2$, $\overline{\boldsymbol{X}}_3$, $\overline{\boldsymbol{X}}_4$, which are characterized by progressively coarser spatial resolutions, with spatial sizes ranging from $\frac{W}{8}\times\frac{H}{8}$ to $\frac{W}{32}\times\frac{H}{32}$. To enable effective fusion, it is essential to unify the spatial dimensions of these multi-scale features. This is achieved using a simple up-sampling block, denoted as f_{up}, which standardizes all features to the same spatial size of $\frac{W}{4}\times\frac{H}{4}$. The up-sampling block f_{up} consists of two vanilla convolution layers followed by a transpose convolution layer. This architecture allows for a two-fold increase in the spatial dimensions of the input feature while simultaneously reducing its channel dimension by half. The transformation process for the multi-scale features is as: 1) The refined feature $\overline{\boldsymbol{X}}_4$, with an initial spatial size of $\frac{W}{32}\times\frac{H}{32}$, undergoes the up-sampling block three times to achieve the target spatial size of $\frac{W}{4}\times\frac{H}{4}$; 2) $\overline{\boldsymbol{X}}_3$, with an initial spatial size of $\frac{W}{16}\times\frac{H}{16}$, is processed through f_{up} twice to reach $\frac{W}{4}\times\frac{H}{4}$; 3) $\overline{\boldsymbol{X}}_2$, initially sized $\frac{W}{8}\times\frac{H}{8}$, passes through the f_{up} once to achieve the desired spatial resolution. These transformations can be formally expressed as follows:

$$\overset{Tran}{4} = f_{up}^3(\overline{\boldsymbol{X}}_4),\ \boldsymbol{X}_3^{Tran} = f_{up}^2(\overline{\boldsymbol{X}}_3),\quad \overset{Tran}{2} = f_{up}^1(\overline{\boldsymbol{X}}_2) \quad (9)$$

where f_{up}^n denotes the application of the up-sampling block n times. The resulting transformed features $\boldsymbol{X}_2^{Tran}$, $\overset{Tran}{3}$, and $\boldsymbol{X}_4^{Tran}$ share the same spatial size ($\frac{W}{4}\times\frac{H}{4}$).

Following the transformation of multi-scale features to a unified spatial resolution, we employ channel concatenation to integrate these features into a single comprehensive representation. This concatenated feature,

which aggregates information from multiple scales, serves as the foundation for subsequent predictions. Finally, we apply a pointwise convolution layer to reduce the channel dimensionality of the concatenated feature, ensuring computational efficiency while retaining essential information. The process of multi-scale feature fusion and dimensionality reduction can be mathematically expressed as:

$$= f_{cat}(\boldsymbol{X}_2^{Tran}, \quad {}_3^{Tran}, \boldsymbol{X}_4^{Tran}) \tag{10}$$

$$\widetilde{\boldsymbol{X}} = f_{pc}(\overline{\boldsymbol{X}}) \tag{11}$$

where f_{cat} represents the channel concatenation operation while f_{pc} denotes the pointwise convolution operation that transforms $\overline{\boldsymbol{X}}$ into $\widetilde{\boldsymbol{X}}$. The detailed structure of the MSFM is shown in Fig. 1(d).

The resulting fused feature $\widetilde{\boldsymbol{X}}$ is subsequently utilized for generating wheat head detection outputs, i.e., the heat maps and bounding box predictions.

IV. Experimental Result

A. Experimental Settings

We evaluate the proposed method on the Global Wheat Head Detection (GWHD) dataset, a benchmark dataset released in 2020 [56]. The GWHD dataset comprises 4700 high-resolution RGB images collected from various countries worldwide, containing approximately 190,000 annotated wheat head instances.

For training and evaluation, the dataset was split randomly into two subsets: 80% of the images were selected as the training set, while the remaining 20% were reserved for testing purposes. To ensure compatibility with the detection model, the input images were resized to a spatial resolution of 512×512 pixels. These resized images were processed by the detection model, which predicted a center heat map and the object size on a feature map with a downsampled spatial resolution of 128×128.

We implemented our proposed detection model using the PyTorch framework on the hardware environment: NVIDIA RTX 3070 GPU (11 GB of VRAM). The software depends on Python 3.8 and Cuda 11.8. The detection models were trained for 300 epochs using the Adam optimization algorithm with a learning rate of 1×10^{-4}.

B. Evaluation Metrics

To evaluate the performance of the proposed detection model, we employed the Average Precision (AP) metric, a standard measure widely used in object detection tasks and defined for the COCO dataset evaluation protocol. The AP metric quantifies the model's overall detection efficacy by computing the area under the Precision–Recall curve (PR curve). This area represents the model's performance across varying confidence thresholds, offering a comprehensive assessment of its accuracy throughout the dataset.

The precision–recall curve captures the relationship between precision and recall, two critical measures in object detection. Therein, precision refers to the proportion of correctly identified wheat heads (True Positives, TP) out of all detections made by the model, including incorrect ones (False Positives, FP). It evaluates the model's reliability in making positive predictions. Whilst recall represents the proportion of actual wheat heads correctly identified by the model (true positives) out of the total number of wheat spikes present in the dataset, including those missed by the model (False Negatives, FN). It assesses the model's ability to identify all relevant objects. The formulas for these calculations are as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{12}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{13}$$

In object detection, the *AP* metric is derived by integrating the precision values at different levels of recall, summarizing the model's performance in a single scalar value, which is computed as:

$$AP = \int_0^1 \text{Precision}(Recall) d(rRecall) \tag{14}$$

In this study, we adopted mAP_{50} as the primary evaluation metric. AP50 measures the average precision when the Intersection over Union (IoU) threshold is set to 0.5. An IoU threshold of 0.5 signifies that a detected bounding box is considered a true positive if it overlaps with the ground truth bounding box by at least 50% of their union. This metric reflects the model's ability to accurately localize and classify wheat spikes within the dataset. Furthermore, we adjust IoU threshold to higher values (0.6, 0.7 and 0.75) to rigorously evaluate and compare the precise localization performance of various models. This adjustment serves as a critical component of the ablation study, enabling a more in-depth analysis of how the proposed components influence the models' ability to accurately delineate object boundaries.

C. Comparison with the Evaluation Metrics

In this study, we present a thorough evaluation of the CenterFormer model for the task of wheat head detection, comparing its performance against ten leading object detection algorithms that represent the state of the art in the field. The comparative analysis, summarized in Table I, encompasses a wide range of detection paradigms to ensure a comprehensive assessment. These include two-stage anchor-based frameworks such as Faster R-CNN [21] and Mask R-CNN [57], which are known for their region proposal mechanisms; one-stage anchor-based methods, including SSD [27], YOLOv3-v5 [24], and EfficientDet [58], which offer faster inference through direct bounding box regression; the anchor-free models like CenterNet [29] and an enhanced version of CenterNet integrated with Multi-Scale Feature Fusion (MSFF) [59]; two Transformer-based models: WheatFormer [53] and CETR [54]. In our experiments, we maintained the original input resolutions as reported in the respective papers to ensure a fair and consistent comparison with their officially published results. The hierarchical SCA-Transformer backbone of our CenterFormer is configured

as ({1,1,9,1}, C=96). Moreover, since the source code for CETR [59] has not been released, we reimplemented CETR based on the description in [59], integrating ResNet18 for multi-scale feature extraction, Feature Pyramid Networks (FPN) [24] for multi-scale feature fusion, and ViT-L/16 for feature refinement, instead of ViT-H/16 as originally used in [59].

TABLE I. COMPARED DETECTION RESULTS WITH THE STATE-OF-THE-ART DEEP MODELS

Methods	mAP_{50}	Input Resolution	FLOPS (G)	Param. (M)
FasterRCNN [10]	0.540	600×600	162.4	42.1
MaskRCNN [57]	0.659	600×600	170.1	74.7
YOLOv3 [15]	0.581	608×608	59.1	59.1
YOLOv4 [16]	0.639	608×608	53.8	63.9
YOLOv5 [17]	0.667	608×608	45.3	21.2
SSD300 [12]	0.649	300×300	35.2	36.1
SSD512 [12]	0.648	512×512	99.5	36.1
EfficientDet [58]	0.701	512×512	20.7	35.3
CenterNet [26]	0.689	512×512	70.2	16.6
CenterNet+MSFF [59]	0.789	512×512	179.5	56.5
WheatFormer [53]	0.725	512×512	99.7	60.1
CETR [54]	0.804	512×512	261.6	220.3
Ours (Base)	0.842	512×512	82.9	47.1

To quantitatively evaluate detection performance, we employed mAP_{50} to measure the trade-off between precision and recall. Our proposed CenterFormer model achieved a noteworthy mAP_{50} score of 0.837, surpassing all other evaluated models. This result underscores the model's advanced detection capabilities, particularly in accurately localizing wheat heads across a variety of complex scenarios, including variations in density, scale, and environmental conditions. Specifically, compared to the recently proposed CenterNet with ViT refinement (CETR) [59], our proposed CenterFormer achieves about a 4% improvement in mAP_{50}, while also having fewer parameters and lower FLOPs. These findings highlight the potential of CenterFormer as a robust and efficient solution for wheat head detection, with implications for advancing precision agriculture practices and automated crop monitoring systems. Fig. 2 manifests the visualization of detection results using our proposed method and six state-of-the-art detection models on three representative samples.

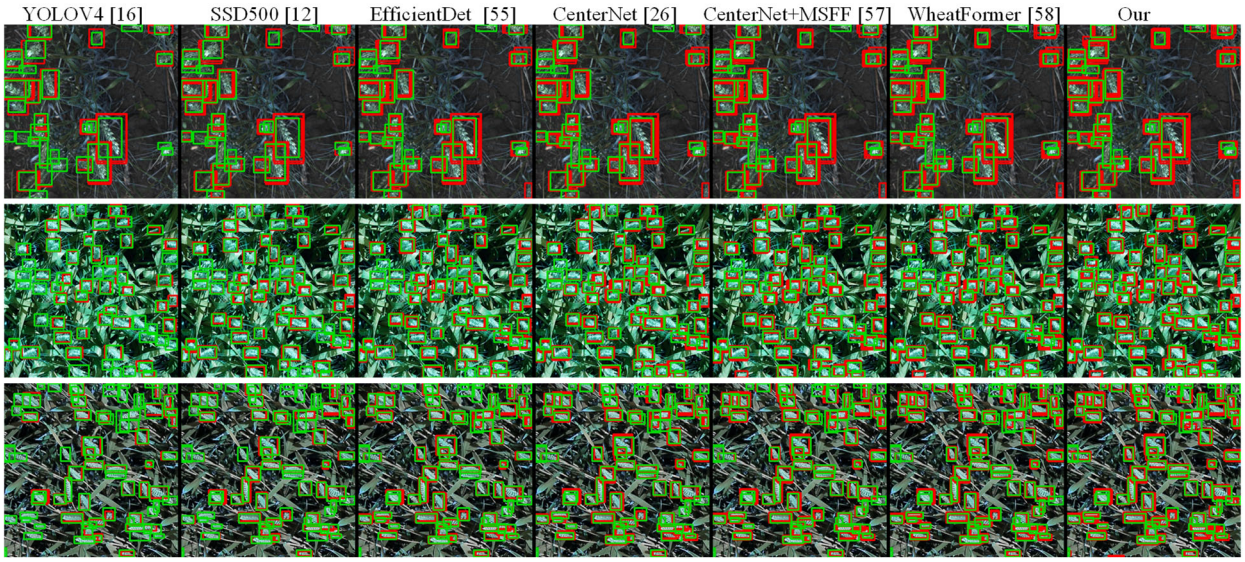


Fig. 2. Visualization of detection results using our proposed method and six state-of-the-art detection models. (Green: Ground-truth; Red: Detection).

Furthermore, to verify potential bias in the dataset split, we divided the GWHD dataset into five groups and employed cross-validation to evaluate the detection performance across all groups. We achieved an average mAP_{50} of 0.844 with a standard deviation of 0.016 over the five runs, indicating that the detection performance is relatively stable across different dataset splits.

Despite the great performance improvements achieved with our proposed CenterFormer, some wheat heads with unclear appearances or severe occlusion remain difficult to detect correctly. Fig. 3 presents examples of both successful and failed detection cases. In successful cases, the wheat heads are relatively distinct from the background and have clear visual boundaries, even under moderate variations in size, orientation, or lighting conditions. CenterFormer effectively captures the spatial patterns and distinguishes the wheat heads despite minor occlusions or overlaps. In the failed cases, detection errors mainly occur due to severe occlusions, or wheat heads with blurred or indistinct appearances. In such challenging conditions, the model struggles to differentiate wheat heads from surrounding noise, leading to missed detections.

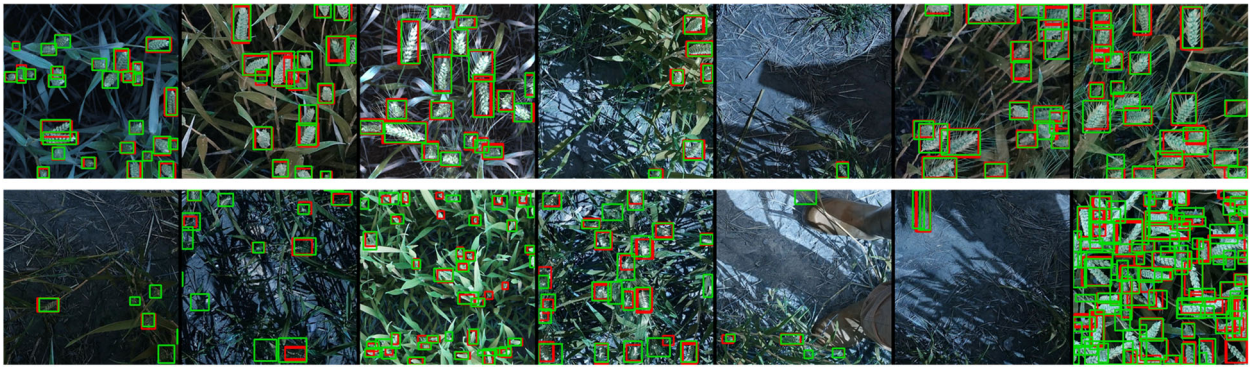


Fig. 3. Examples of successful and failed cases (Green: Ground-truth; Red: Detection). Top: successful examples; Bottom: examples with different degrees of failure detection.

D. Hyperparameter Sensitivity Analysis in SCA Backbone

Following the experimental setup in [49], we fixed the patch size and the channel number per head, obtaining $P = 49$ and $C_h = 32$ in the spatial window attention while set the channel number per group as $C_g = 32$ in the channel group attention for the micro SCA architecture. However, we varied the number of SCA-Transformer blocks and the overall channel dimensions per stage using three configurations: $(\{1, 1, 3, 1\}, C = 96)$, $(\{1, 1, 9, 1\}, C = 96)$ and $(\{1, 1, 9, 1\}, C = 128)$. With the channel number per head ($C_h = 32$) and per group ($C_g = 32$) fixed, these configurations result in different numbers of attention heads and groups across the network stages. Accordingly, the three settings correspond to three model variants: Small-SCA, Base-SCA, and Large-SCA. We conducted ablation experiments with these three variants, and the results are summarized in Table II, highlighting their impact on detection performance (AP), computational complexity (FLOPs), and model size (number of parameters). This analysis illustrates the trade-offs between accuracy and efficiency, providing empirical justification for the final model configuration adopted in *CenterFormer*.

E. Ablation Experiments

As previously introduce, the primary objective of this study is to explore the feasibility of substituting convolutional-based structures with transformer-based blocks in deep learning architectures. To evaluate the efficacy of the learned feature representations at various stages within the network, we utilized both the conventional ResNet50 model and our proposed SCA-Transformer backbone. Features were extracted from these architectures at the second to fourth stages (blocks) of processing, denoted as S2, S3 and S4, respectively, and their impact on detection performance was assessed. The evaluation was conducted using IoU thresholds of 0.5, 0.6, 0.7, and 0.75. The comparative detection results are presented in Table III(a). A detailed examination of Table III(a) reveals that the SCA-Transformer backbone demonstrates a remarkable improvement in detection performance across multiple scales of feature maps when compared to the ResNet50 backbone. These results underscore the potential of our proposed SCA-Trans backbones in enhancing feature representation and detection accuracy.

TABLE II. ABLATION STUDY WITH DIFFERENT CONFIGURATIONS OF THE SCA-TRANSFORMER BACKBONE

Backbones	mAP_{50}	FLOPS (G)	Param. (M)
Small-SCA	0.834	66.1	30.7
Base-SCA	0.842	82.9	47.1
Large-SCA	0.837	142.4	83.2

TABLE III. ABLATION EXPERIMENTS

Performance Comparisons	Methods	mAP_{50}	mAP_{60}	mAP_{70}	mAP_{75}
The used feature maps at different stages	ResNet50-S2	0.483	0.402	0.213	0.147
	SCA-Trans-S2	0.742	0.599	0.355	0.224
	ResNet50-S3	0.664	0.515	0.293	0.176
	SCA-Trans-S3	0.829	0.705	0.454	0.301
	ResNet50-S4	0.689	0.561	0.321	0.193
	SCA-Trans-S4	0.828	0.703	0.449	0.296
The MSFM and CSAM	ResNet50+MSFM34	0.775	0.646	0.392	0.245
	SCA-Trans+MSFM34	0.822	0.707	0.464	0.312
	ResNet50+MSFM234	0.761	0.643	0.408	0.255
	SCA-Trans+MSFM234	0.826	0.715	0.464	0.315
	SCA-Trans+MSFM234+CSAM	**0.842**	**0.731**	**0.478**	**0.322**

Subsequently, we assess the effectiveness of the proposed Multi-Scale Fusion Module (MSFM) by combining the features from different stages of the network. Specifically, we aggregated the feature maps from stages 3 and 4, as well as from stages 2, 3, and 4, for both the ResNet50 and our SCA-Transformer backbones. The comparative results of these aggregations are presented in Table III(b), where the performance of each

method is evaluated across these different feature combinations. An analysis of the results presented in Table III(b) reveals that the aggregation of features usually enhances detection performance especially for the mAP with high IoU values. This improvement underscores the effectiveness of combining feature maps from multiple stages of the network. Furthermore, Table III(b) also highlights the impact of incorporating the CSFM for feature refinement, illustrating its contribution to further enhancing the detection accuracy. To evaluate the effect of the CSFM module, we visualize the feature maps: $\boldsymbol{X}_2$, $\boldsymbol{X}_3$ and $\boldsymbol{X}_4$ before applying CSFM, and $\bar{\boldsymbol{X}}_2$, $\bar{\boldsymbol{X}}_3$ and $\bar{\boldsymbol{X}}_4$ after applying CSFM, as shown in Fig. 4. These visualizations clearly highlight how CSAM enhances important regions while suppressing less informative areas.

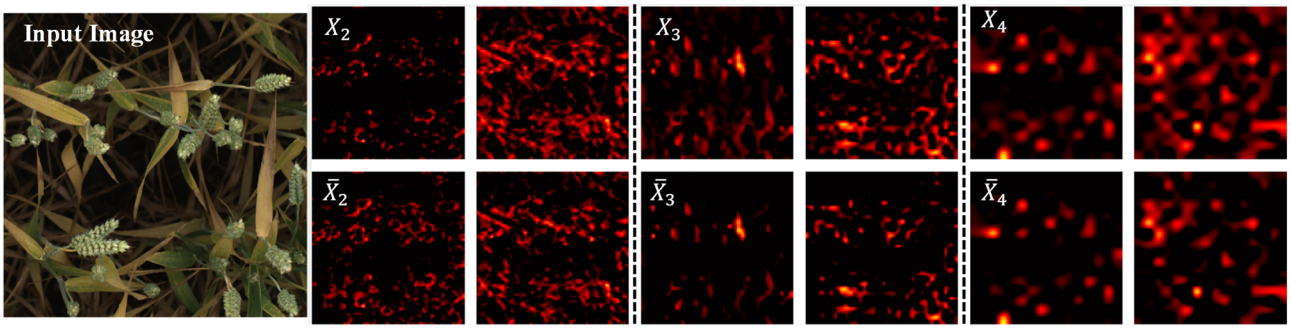


Fig. 4. Visualization of the feature maps before/after CSFM.

F. Discussions

While our proposed CenterFormer demonstrates strong performance on wheat head detection tasks, we acknowledge that computational efficiency (approximately twice the computational cost of YOLOv5) and deployability still require improvement, particularly for real-world applications on resource-constrained platforms such as edge devices. Our current implementation employs a hierarchical Transformer backbone combined with multi-scale feature processing, which, while effective, introduces a non-negligible computational overhead. In our experiments, the Small-SCA configuration achieves comparable performance to the base setup while reducing computational cost by approximately 25%, suggesting that a smaller backbone may lead to more deployable detection models in real scenarios. To address deployment challenges, several potential optimization strategies such as model compression/acceleration, efficient backbone alternatives (e.g. MobileViT, EfficientFormer), edge-specific adaptation are anticipated in the future work.

Although our work primarily focuses on wheat head detection, the architecture and principles of our approach hold significant promise for extension to a variety of other agricultural tasks. First, the ability to identify fine-grained spatial patterns makes our method particularly well-suited for plant disease detection, where subtle visual symptoms must be accurately recognized across varying scales and growth stages. Similarly, the multi-scale feature extraction and fusion strategies are applicable to crop yield estimation, where large-scale and heterogeneous spatial information needs to be processed efficiently. Moreover, the flexible and modular design of our approach allows for straightforward adaptation to other remote sensing-based agricultural tasks through fine-tuning on the relevant datasets.

In future work, we plan to explore these extensions and systematically evaluate the generalization ability of our method across diverse agricultural scenarios. Such efforts will further highlight the versatility and practical value of our framework in advancing precision agriculture.

V. Conclusion

In this study, we have proposed a novel approach, CenterFormer, for automatic wheat head detection that effectively addresses the challenges posed by dense object annotations. By treating the wheat head as a single point and leveraging the long-range dependency modeling capabilities of the Transformer architecture, our method can learn multi-scale features crucial for accurate head prediction. The hierarchical Transformer backbone with self-attention mechanisms in both spatial and channel domains, combined with window-based and group-wise attention strategies, ensures efficient feature extraction while maintaining linear complexity. Moreover, the introduction of a fusion block for multi-scale feature integration enhances both detailed spatial information and semantic context, significantly improving wheat head detection performance. Our extensive experiments on the Global Wheat Head Detection (GWHD) dataset demonstrate that CenterFormer outperforms existing state-of-the-art detection models, highlighting its effectiveness and potential for real-world applications in precision agriculture. Future work will focus on further improving the scalability and robustness of the model across different environments and crop types.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, methodology and software, E.H. and X.H.; validation, E.H.; formal analysis and investigation, writing—original draft preparation, review and editing, X.L.; writing—review and editing, E.H. and X.H.; project

administration and funding acquisition, X.H. all authors had approved the final version.

FUNDING

This research was supported by a research grant from the Okawa Foundation for Information and Telecommunications Foundation.

REFERENCES

[1] F. Catherine, F. Klaus, X. Nayer, J. Rogers, and K. Eversole, "Slicing the wheat genome," *Science*, vol. 345, pp. 285–285, 2014.

[2] I. Delabre, L. O. Rodriguez, J. M. Smallwood *et al.*, "Actions on sustainable food production and consumption for the post-2020 global biodiversity framework," *Sci. Adv.*, vol. 7, eabc8259, 2021.

[3] A. Watson, S. Ghosh, M. J. Williams, W. S. Cuddy *et al.*, "Speed breeding is a powerful tool to accelerate crop research and breeding," *Nat. Plants*, vol. 4, pp. 23–29, 2018.

[4] X. Zhang, H. Jia, T. Li, J. Wu *et al.*, "TaCol-B5 modifies spike architecture and enhances grain yield in wheat," *Science*, vol. 376, pp. 180–183, 2018.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. International Conference on Learning Representations (ICLR)*, 2015.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[7] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2261–2269.

[8] Y. Li, K. He, J. Sun *et al.*, "R-FCN: Object detection via region-based fully convolutional networks," *Advances in Neural Information Processing Systems (NIPS)*, pp. 379–387, 2016.

[9] R. Girshick, "Fast R-CNN," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems (NIPS)*, pp. 91–99, 2015.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, pp. 1137–1149, 2017.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. European Conference on Computer Vision (ECCV)*, 2016, pp. 21–37.

[13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, real-time object detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[14] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," arXiv preprint, arXiv:1612.08242, 2016.

[15] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv preprint, arXiv:1804.02767, 2018.

[16] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv preprint, arXiv:2004.10934, 2020.

[17] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021.

[18] M. Everingham, S. M. A. Eslami, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015.

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. European Conference on Computer Vision (ECCV)*, 2014, pp. 740–755.

[20] B. Gong, D. Ergu, Y. Cai, and B. Ma, "A method for wheat head detection based on YOLOv4," *Research Square*, 2020.

[21] B. Gong, D. Ergu, Y. Cai, and B. Ma, "Real-time detection for wheat head applying deep neural network," *Sensors*, vol. 21, no. 1, p. 191, 2021.

[22] S. Khaki, N. Safaei, H. Pham, and L. Wang, "A lightweight convolutional neural network for high-throughput image-based wheat head detection and counting," arXiv preprint, arXiv:2103.09408, 2021.

[23] A. Thakur, S. Singh, N. Goyal, and K. O. Gupta, "A comparative analysis on the existing techniques of wheat spike detection," in *Proc. 2nd International Conference for Emerging Technology*, 2021.

[24] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 936–944.

[25] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. European Conference on Computer Vision (ECCV)*, 2018, pp. 765–781.

[26] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," arXiv preprint, arXiv:1904.07850, 2019.

[27] X. Zhou, J. Zhuo, and P. Krähenbühl, "Bottom-up object detection by grouping extreme and center points," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 850–859.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 6000–6010, 2017.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn *et al.*, "An image is worth 16×16 Words: Transformers for image recognition at scale," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[30] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9992–10002.

[31] Z. Liu, H. Hu, Y. Lin *et al.*, "Swin transformer V2: Scaling up capacity and resolution," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11999–12009.

[32] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," arXiv preprint, arXiv:2012.12877, 2020.

[33] C. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," arXiv preprint, arXiv:2103.14899, 2021.

[34] Y. Li, K. Zhang, J. Cao, R. Timofte, and L. V. Gool, "LocalViT: Bringing locality to vision transformers," arXiv preprint, arXiv:2104.05707, 2021.

[35] M. M. Hasan, J. P. Chopin, H. Laga, and S. J. Miklavcic, "Detection and analysis of wheat spikes using convolutional neural networks," *Plant Methods*, vol. 14, p. 100, 2018.

[36] S. Madec, X. Jin, H. Lu, B. De Solan, S. Liu, F. Duyme, E. Heritier, and F. Baret, "Ear density estimation from high resolution RGB imagery using deep learning technique," *Agric. For. Meteorol.*, vol. 264, pp. 225–234, 2019.

[37] B. Yang, Z. Gao, Y. Gao, and Y. Zhu, "Rapid detection and counting of wheat ears in the field using YOLOv4 with attention module," *Agronomy*, vol. 11, no. 6, p. 1202, 2021.

[38] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[39] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[40] J. Sun, K. Yang, C. Chen, J. Shen, Y. Yang, X. Wu, and T. Norton, "Wheat head counting in the wild by an augmented feature pyramid networks-based convolutional neural network," *Comput. Electron. Agric.*, vol. 193, 106705, 2022.

[41] J. Ye, Z. Yu, Y. Wang, D. Lu, and H. Zhou, "WheatLFANet: In-field detection and counting of wheat heads with high-real-time global regression network," *Plant Methods*, vol. 19, p. 103, 2023.

[42] J. Yan, J. Zhao, Y. Cai *et al.*, "Improving multi-scale detection layers in the deep learning network for wheat spike detection based on interpretive analysis," *Plant Methods*, vol. 19, p. 46, 2023.

[43] J. Zhao, Y. Cai, S. Wang *et al.*, "Small and oriented wheat spike detection at the filling and maturity stages based on WheatNet," *Plant Phenomics*, vol. 5, p. 0109, 2023.
[44] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," arXiv preprint arXiv:2010.04159, 2021.
[45] Z. L. Yang, W. H. Yang, J. Z. Yi, and R. Liu, "WH-DETR: An efficient network architecture for wheat spike detection in complex backgrounds," *Agriculture*, vol. 14, p. 961, 2024.
[46] S. An, S. Park, G. Kim, J. Baek, B. Lee, and S. Kim, "Context enhanced transformer for single image object detection in video data," in *Proc. the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence,* 2023, pp. 682–690.
[47] S. Zheng, J. Lu, H. Zhao *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6881–6890.
[48] H. Wang, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "Max-Deeplab: End-to-end panoptic segmentation with mask transformers," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 5459–5470.
[49] M. Ding, B. Xiao, N. Codella, P. Luo, J. Wang, and L. Yuan, "DaViT: Dual attention vision transformer," arXiv preprint, arXiv:2204.03645, 2022.
[50] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, "DAB-DETR: Dynamic anchor boxes are better queries for DETR," in *Proc. International Conference on Learning Representations (ICLR)*, 2022.
[51] D. Meng, X. Chen, Z. Fan, G. Zeng, H. Li, Y. Yuan, L. Sun, and J. Wang, "Conditional DETR for fast training convergence," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 3631–3640.
[52] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection," arXiv preprint, arXiv:2203.03605, 2022.
[53] Q. Zhou, Z. Huang, S. Zheng, L. Jiao, L. Wang, and R. Wang, "A wheat spike detection method based on transformer," *Front. Plant Sci.*, vol. 13, 1023924, 2022.
[54] K. G. Suma, G. Sunitha, R. Karnati, E. R. Aruna, K. Anvesh, N. Kale, and P. K. Kishore, "CETR: CenterNet-vision transformer model for wheat head detection," *J. Auton. Intell.*, vol. 7, no. 3, 100265, 2024.
[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," arXiv preprint, arXiv:1512.03385, 2015.
[56] E. David, S. Madec, P. Sadeghi-Tehran, H. Aasen, B. Zheng *et al.*, "Global Wheat Head Detection (GWHD) dataset: A large and diverse dataset of high-resolution RGB-labelled images to develop and benchmark wheat head detection methods," *Plant Phenomics*, vol. 2020, 3521852, 2020.
[57] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.
[58] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," arXiv preprint, arXiv:1911.09070, 2019.
[59] S. Harada and X.-H. Han, "Coarse-to-fine pyramid feature mining for wheat head detection," in *Proc. IEEE International Conference on Image Processing (ICIP)*, 2023, pp. 1350–1354.