Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.063287

ARTICLE

TransSSA: Invariant Cue Perceptual Feature Focused Learning for Dynamic Fruit Target Detection

Jianyin Tang, Zhenglin Yu* and Changshun Shao

School of Mechanical and Electrical Engineering, Changchun University of Science and Technology, Changchun, 130022, China

*Corresponding Author: Zhenglin Yu. Email: yuzhenglin@cust.edu.cn

Received: 10 January 2025; Accepted: 20 February 2025; Published: 16 April 2025

ABSTRACT: In the field of automated fruit harvesting, precise and efficient fruit target recognition and localization play a pivotal role in enhancing the efficiency of harvesting robots. However, this domain faces two core challenges: firstly, the dynamic nature of the automatic picking process requires fruit target detection algorithms to adapt to multi-view characteristics, ensuring effective recognition of the same fruit from different perspectives. Secondly, fruits in natural environments often suffer from interference factors such as overlapping, occlusion, and illumination fluctuations, which increase the difficulty of image capture and recognition. To address these challenges, this study conducted an in-depth analysis of the key features in fruit recognition and discovered that the stem, body, and base serve as constant and core information in fruit identification, exhibiting long-term dependent semantic relationships during the recognition process. These invariant features provide a stable foundation for dynamic fruit recognition, contributing to improved recognition accuracy and robustness. Specifically, the morphology and position of the stem, body, and base are relatively fixed, and the effective extraction of these features plays a crucial role in fruit recognition. This paper proposes a novel model, TransSSA, and designs two innovative modules to effectively extract fruit image features. The Self-Attention Core Feature Extraction (SAF) module integrates YOLOV8 and Swin Transformer as backbone networks and introduces the Shuffle Attention self-attention mechanism, significantly enhancing the ability to extract core features. This module focuses on constant features such as the stem, body, and base, ensuring accurate fruit recognition in different environments. On the other hand, the Squeeze and Excitation Aggregation (SAE) module combines the network's ability to capture channel patterns with global knowledge, further optimizing the extraction of effective features. Additionally, to improve detection accuracy, this study modifies the regression loss function to EIOU. To validate the effectiveness of the TransSSA model, this study conducted extensive visualization analysis to support the interpretability of the SAF and SAE modules. Experimental results demonstrate that TransSSA achieves a performance of 91.3% on a tomato dataset, fully proving its innovative capabilities. Through this research, we provide a more effective solution for using fruit harvesting robots in complex environments.

KEYWORDS: Fruit recognition; invariant features; TransSSA model; swin transformer; self-attention mechanism

1 Introduction

Object detection development comprises two main stages: the traditional stage and the deep learning stage based on convolutional neural networks. After the advent of convolutional networks, models like AlexNet [1], VGG [2], GoogLeNet [3–5], and ResNet [6] have drawn wide attention. Currently, research on object detection emphasizes static object detection with significant achievements.

The world increasingly values agricultural efficiency and sustainable development, making automated harvesting technology crucial in modern agriculture. Fruit recognition, a key part of automated harvesting

systems, aims to enhance harvesting efficiency and quality, ensure timely harvests, and cut losses. With rising global fruit production, along with increasing labor costs and manpower shortages, the demand for efficient fruit detection technologies grows [7]. Fruit detection is a challenging dynamic object detection problem. In natural environments, issues like overlapping, occlusion, and illumination changes during photography impede accurate identification.

In recent years, deep learning technology has advanced rapidly. Fruit object detection has thus become a research focus, with methods typically divided into two categories: the Region Detection Method (RDM), which extracts and analyzes regions of interest (RoIs) from input images, uses deep convolutional neural networks (DCNNs) for feature extraction, and locates fruits with the Region Proposal Network (RPN), such as in Faster R-CNN and Mask R-CNN algorithms. After integrating the RPN, Faster R-CNN's detection speed and accuracy improved notably, outperforming traditional techniques in complex-background fruit detection [8]. Mask R-CNN extends Faster R-CNN, enhancing segmentation ability to generate fruit bounding boxes and masks simultaneously, improving recognition accuracy [9].

Compared to the RDM, the Image Detection Method (IDM) must directly analyze entire visual images and handle vast visual data to extract discriminative features. Convolutional neural networks (CNNs) and fully convolutional networks (FCNs) have made remarkable progress here. The YOLO algorithm family, turning object detection into a regression task for real-time recognition, shows excellent speed and accuracy in fruit recognition, fitting agricultural scenarios needing quick responses [10,11]. The Single Shot Multi Box Detector (SSD) detects across different scale feature maps to target fruits of various sizes, exhibiting outstanding detection capabilities in multiple datasets [12].

Despite their remarkable performance in fruit object detection, the RDM and IDM face challenges. The RDM is susceptible to false positives in complex backgrounds, while the IDM is fast but has accuracy limitations. Hence, developing novel network architectures and optimizing algorithms are core research directions.

1.1 Challenges

The intrinsic properties of outdoor fruit images impede high accuracy fruit target detection, mainly in three aspects. Firstly, for dynamic target detection, as the harvesting robot moves in the orchard, the camera perspective varies, causing the same fruit's morphological changes and complicating detection and localization, like the apples shown from different perspectives in the left picture of Fig. 1. Secondly, fruit overlapping and occlusion are common during growth, with leaves partially hiding fruits. Traditional algorithms often make mistakes in such cases, so effective strategies are urgently needed, as seen in the middle of Fig. 1. Thirdly, illumination variability matters. Lighting changes can alter fruit color, brightness and contrast in images, reducing detection precision. In Fig. 1 right, improper light decreases apple image contrast and the difference from the background, leading to recognition difficulties and false detections, especially in complex backgrounds.

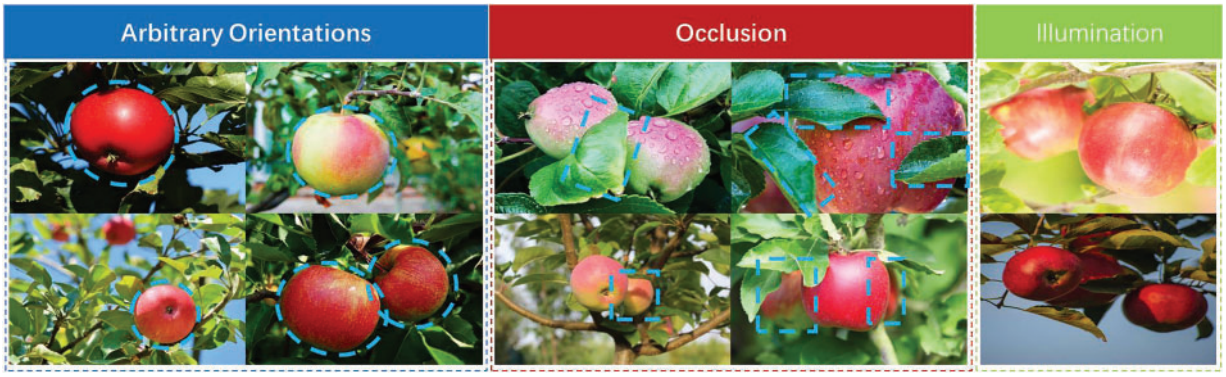

Figure 1: The fruit recognition is faced with the challenges of arbitrary orientation, occlusion and illumination, which can cause partial information loss in fruit images. On the left, the fruit image is affected in any direction. The middle fruit image is affected by the occlusion; On the right, the fruit image is affected by light

1.2 Observation and Motivation

When carefully examining fruit images, it became clear that these images have a number of distinctive features. Due to the variations in shooting angle, fruit images initially exhibit a variety of shapes, including those viewed from base, parallel, and top perspectives. Each of these morphological variants has unique characteristics. Fortunately, there are invariant characteristics in these forms that can be used to identify and locate fruits. Additionally, fruit images from natural environments often suffer from challenges such as overlap, occlusion, and variations in lighting. These factors make the extraction of fruit features more complex and mean that traditional image processing techniques are often inadequate in dealing with these problems.

Finding I: The invariable evidence of congeneric fruits. Fruit identification goes beyond just visual appearance. The essential characteristics of a category must be differentiated from other characteristics. Once an algorithm recognizes these crucial characteristics, it makes it easier to recognize fruits within the same category. Fig. 2 shows an illustrative example. The figure describes four different perspectives of an apple, with each perspective containing a variety of attributes. However, when these attributes are analyzed together, they can result in misleading information that is detrimental to effective apple identification. By extracting the core features and identifying the long-term semantic dependency relationships between apple fruits—for example, the connections between the stem, body and base of the fruit such adverse effects can be mitigated. As highlighted by the colored bounding boxes in Fig. 2, the stem, body, and base of the apple are the predominant features that contribute to its accurate recognition.

Finding II: The images of fruits from natural environments often face challenges such as overlay, occlusion, and variations in lighting. These factors make the extraction of fruit characteristics a more complex undertaking, with traditional image processing techniques often proving inadequate in dealing with this complexity. Consequently, the art of isolating effective fruit characteristics within the labyrinth of the natural environment has become central to the pursuit of efficient automated harvesting. The above discovery can be summarized as a quest to identify invariant core features within a class of fruit images while increasing the effectiveness of feature extraction. In particular, we argue that detecting these core features which are insensitive to variations from different perspectives of fruit images within the same category is paramount in conjunction with improving the ability to extract relevant features. Consequently, the judicious application of these findings plays a crucial role in improving the precision of dynamic fruit target detection. Our impetus

is to develop an innovative method that can efficiently extract features from fruit images to improve the performance of dynamic fruit target detection.

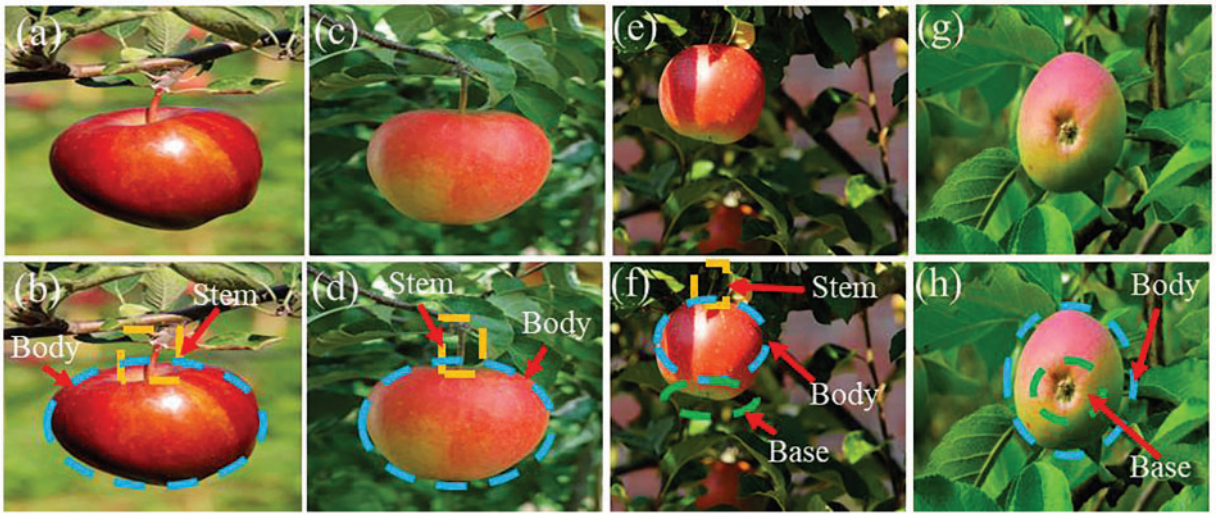

Figure 2: A series of apple images taken from different perspectives. Panels (a, c, e, g) represent photographs taken from four different angles, while (b, d, f, h) show the correlation between core identification represent features of the fruit, namely the stem, the body and the base

To achieve this goal, we have developed a novel model that we call TransSSA, as proposed in this paper. We have developed two novel modules: the Self-Attention Core Feature Extraction (SAF) module and the Squeeze and Excitation Aggregation (SAE) module. The SAF module combines YOLOV8 and Swin Transformer as a backbone network, while integrating the Self-Attention mechanism known as Shuffle Attention (SA) to support core feature extraction. Through this approach, we are better able to capture the characteristic features of fruits from different perspectives and overcome the challenges posed by morphological changes in dynamic recognition. The SAE module, on the other hand, integrates the network's ability to capture channel patterns and global knowledge. This module improves the extraction of effective fruit features in overlap and occlusion scenarios, thereby refining the accuracy of detection. In parallel, we modified the regression loss function in EIOU to increase the detection accuracy.

1.3 Contributions

Our investigation addressed the invariant features of dynamic fruit target recognition, driven by a dual motivation: first, the pursuit of extracting intrinsic features that define similar fruits; secondly, improving the ability to effectively extract salient features of fruits. Leveraging the insights gained from our observations, we introduced an innovative TransSSA model designed to learn the invariant cues and essential attributes of fruit images. The main contributions of this study are as follows: (1) We developed an efficient TransSSA model to exploit the observations made in the fruit dataset, focusing on the invariant cues present in images of similar fruits. By using the Swin Transformer to explore these relationships, we extracted the core functions required to generate feature maps within the TransSSA model. (2) We proposed two novel modules to exploit multi-scale information and extract the essential features of fruit images while improving the extraction of salient attributes. In particular, the SAF module expanded the ability to extract core features, while the SAE module improved the extraction of effective fruit features in overlap and occlusion scenarios, thereby improving the recognition accuracy. In addition, we changed the regression loss function to EIOU to refine the detection

accuracy. These modifications collectively used the information in fruit images to build an invariant cue-aware deep learning neural network. (3) We conducted experiments on the Fruit dataset and the proposed TransSSA model outperformed several state-of-the-art methods, confirming its effectiveness.

1.4 Organization of This Paper

The rest of the work is structured as follows. Recent work on dynamic fruit target detection is presented in Section 2. The details of our model are explained in Section 3. The experimental results, discussion of two different data sets and extended experiments in Section 4. Finally, we conclude our work in Section 5.

2 Related Work

2.1 Problem Formulation

In general, the goal of fruit detection can be summarized as follows: when an image x is presented with multiple objects along with their corresponding categories y and bounding box information b, the goal is to create a mapping function G that estimates the predicted categories $\hat{y}$ and bounding box $\hat{b}$. The essence of this procedure is to ensure that $\hat{y}$ and $\hat{b}$ represent the actual target information in the image as accurately as possible. In recent years, a variety of network architectures have emerged, including Convolutional Neural Networks (CNNs), Region-based Convolutional Neural Networks (R-CNNs), YOLO (You Only Look Once), and SSD (Single Shot MultiBox Detector) [13]. These deep learning approaches manipulate image features in different ways and use regression and classification mechanisms to predict the category and location of targets. During the network training phase, the model parameters are optimized by minimizing the cross-entropy loss between the predicted categories $\hat{y}$ and the actual categories y, as well as the smooth L_1 loss between the predicted bounding boxes $\hat{b}$ and the true bounding boxes b.

2.2 Dynamic Fruit Target Detection

Numerous studies address the challenge of dynamic fruit target recognition, mainly applying deep learning (DL) techniques divided into three methods.

Regional Proposal Network (RPN) approaches use DL models to generate candidate regions for fruit classification and localization. Ren et al. [14] introduced Faster R-CNN, enhancing object detection speed and accuracy via a regional proposal network, sharing folding features to reduce computation and accurately position fruits. Zhou et al. [15] optimized RPN's loss function to improve small fruit detection.

Feature Pyramid Network (FPN) methods create multi-scale feature maps for fine-grained detection. Lin et al. [16] proposed FPN architecture, integrating features at different scales to boost fruit identification and localization, performing well on various datasets, especially for small fruits and dense scenes. Chen et al. [17] incorporated contextual information into FPN to refine feature representation and recognition accuracy in complex backgrounds.

Self-Attention Mechanism (SA) methods introduce attention to focus on salient features. Vaswani et al. [18] presented Transformer architecture, integrating Self-Attention into DL models for dynamic focus adjustment. Li et al. [19] used Self-Attention to enhance fruit feature expressiveness, improving recognition.

In summary, FPN and SA excel in detecting small targets and complex backgrounds, while RPN's efficiency matters in fast-processing scenarios.

2.3 Yolo and Transformer-Based Image Detection

The Transformer architecture, initially prevalent in natural language processing, deeply influenced computer vision. Carion et al. [20] presented a Transformer-based end-to-end object detection method,

revealing its visual task potential. Zhu et al. [21] proposed a deformable detection transformer and explored variants of end-to-end object detectors. D. Alexey et al. [22] introduced Vision Transformer (ViT), showing it could rival and potentially replace traditional CNNs. Building on this, Alshawabkeh et al. [23] proposed a hybrid approach combining Mask R-CNN and the Vision Transformer (ViT) model for pavement crack detection. Xue et al. [24] combined CNN and Transformer for facial expression recognition.

Despite ViT's achievements, it has drawbacks like computational inefficiency and poor small object detection. ViT divides images into patches, but the quadratic complexity-patch size relationship hinders capturing fine-grained info. Inspired by ViT, Swin Transformer performs calculations in a shift window, leveraging fine-grained representations to handle intensive tasks, and achieves great results in object detection and image segmentation [25–27].

In our study on dynamic fruit target detection, challenges like variable angles, lighting, and fruit occlusion demand perception of image details, where traditional CNNs falter. We introduce Transformer for its strength in mining latent details and semantic relations. Given the need for fast processing in automated harvesting, we integrate YOLOV8 and Swin Transformer as a backbone network, leveraging the Feature Pyramid Network (FPN) method's suitability for speed.

3 Method

3.1 Overview

The process flow of the TransSSA model is shown in Fig. 3. This methodology includes two main components: the SAF module and the SAE module. The first segment, the SAF module, serves as a central framework for feature extraction and is tasked with delineating fine-grained, multi-scale information from fruit images. As part of this research, we merged YOLOV8 and Swin Transformer as a backbone net-work because of their exceptional ability to leverage shift-window schemes to uncover long-distance semantic dependency relationships within fruit images such as stem, body, and base. The network culminates in the extraction of invariant core features, with the backbone further refined to incorporate the Self-Attention mechanism known as Shuffle Attention (SA), thereby expanding the capability of core feature extraction. The second component, the SAE module, represents an innovative detection head. This module is able to improve the extraction of effective fruit features in overlap and occlusion scenarios, thereby increasing the accuracy of detection. Ultimately, we modified the regression loss function in EIOU to improve the detection accuracy.

3.2 SAF Module

The pipeline of the SAF model is presented in Fig. 3. The proposed method includes two parts: a feature extraction backbone and Shuffle Attention. The first part, the feature extraction backbone, is tasked with extracting multi-scale information from fruit images. In this study, the Swin Transformer is adopted as the backbone due to its exceptional performance in uncovering long-dependency semantic relationships within fruit images, such as those related to the fruit stem, body and base, through its shifted window approach. The second part, Shuffle Attention, integrates channel and spatial attention mechanisms by shuffling and reordering input data to compute attention weights. This mechanism enables the model to more accurately focus on critical information within input sequences and feature maps, thereby enhancing its core feature extraction capability.

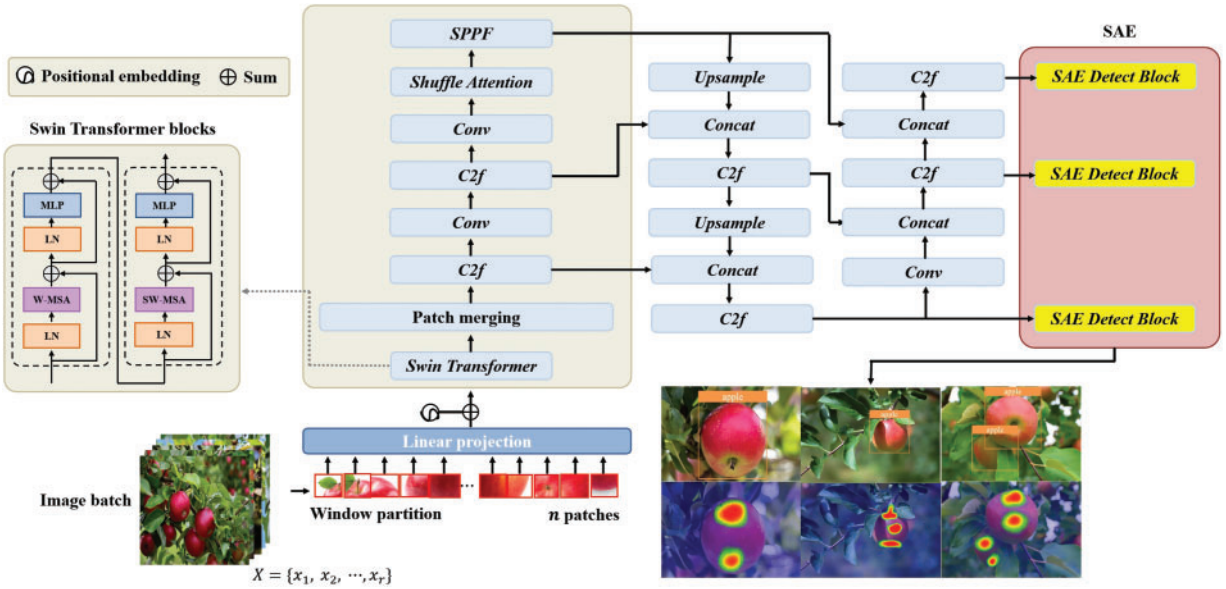

Figure 3: Overview of the TransSSA architecture. First, the image is divided into window partitions and attention is calculated only within the window. Secondly, patch is trans-formed into one-dimensional feature vector by linear projection and position embedding. Third, the SAF module integrates YOLOV8 and Swin Transformer as the backbone network to first follow the Swin Transformer block to utilize the long dependency semantic relationship between feature vectors. At the same time, the Shuffle Attention mechanism is integrated to expand the core feature extraction capability. Fourth, the last stage SAE module as a new detection head combines the ability of the network to capture channel patterns with global knowledge to improve the extraction of effective features. In addition, we change the regression loss function to EIOU to improve the detection accuracy. Finally, the input fruit images are recognized and predicted

3.2.1 Feature Map Generation

The process of converting fruit images into a feature map by extracting salient structural components can be divided into three main phases: First, the fruit image is divided into numerous smaller segments; position embeddings are then applied; and finally, the Transformer blocks that use the DPG strategy are used for processing. After completing these steps, the original input image is converted into a feature map.

Phase I: Segmentation of the image into smaller segments. In this initial phase, the raw input image is first divided into non-overlapping segments, with each segment representing a one-dimensional vector. Specifically, the input image, which initially has size $X \in \mathbb{R}^{H\times W\times C}$, is divided into a grid of smaller segments. The number of these segments can be calculated using the following formula:

$$n = \frac{H}{P_h} \times \frac{W}{P_w}, \tag{1}$$

where n denotes the number of patches, where P_h and P_w represent the height and width of each patch, respectively. Each patch p_i is then flattened into 1D vector of size $P_h \times P_w \times C$ linear projection is then applied to p_i, which is then projected onto $p_i^{'}$. This procedure can be formulated as follows:

$$p_i^{'} = p_i \cdot E, \quad i\varepsilon 1, 2, 3, \cdots, n, \tag{2}$$

where $p_i \in \mathbb{R}^{(P_h\times P_w\times C)}$ denotes the i-th patch, $E \in \mathbb{R}^{(P_h\times P_w\times C)\times d}$ denotes the linear projection and $p_i^{'} \in \mathbb{R}^{d}$ represents the visual vector one d-dimensional projection.

Phase II: Positional embedding is introduced. Since the transformer layer is invariant to permutation of the input patch sequence, position embeddings are important to encode the spatial position and relationships of the patches. In the following phase, the position meaning is embedded. Since the transformer layer remains insensitive to the reordering of the input sequence of patches, the inclusion of position embeddings is essential for encoding the spatial coordinates and interdependencies of these patches.

In particular, these patches are integrated into the patch vectors by adding position embeddings. The formula is as follows:

$$c_0 = \left[p_1^{'}, p_2^{'}, p_3^{'}, \cdots, p_n^{'}\right] + E_{pos}, \tag{3}$$

where $c_0 \in \mathbb{R}_{n\times d}$ denotes a matrix consisting of patch vectors, where "n" represents the number of patches and $E_{pos} \in \mathbb{R}_{n\times d}$ means position embeddings. The type of position embeddings can be selected from a variety of options, including 2D sine embeddings, learnable embeddings, and relative position embeddings.

Phase III: Navigating the Transformer Constructs. After position embedding, the individual patches are systematically processed by a series of M Swin Transformer modules. Each Swin Transformer module is carefully calculated according to the following protocol:

$$\begin{cases} \hat{d}^l = W - MSA\left[LN\left(d^{l-1}\right)\right] + d^{l-1}, \\ d^l = MLP\left[LN\left(\hat{d}^l\right)\right] + \hat{d}^l, \\ \hat{d}^{l+1} = SW - MSA\left[LN\left(d^l\right)\right] + d^l, \\ d^{l+1} = MLP\left[LN\left(\hat{d}^{l+1}\right)\right] + \hat{d}^{l+1}, \end{cases} \tag{4}$$

where in $\hat{d}^l$ and d^l denote the output patch vectors of the (S)W-MSA module and the *MLP* module, respectively, within the l-th Ttransformer block, respectively. *LN* refers to layer normalization. *MLP* signifiesrefers to a cascade of fully connected layers. *W-MSA* stands for window-based multi-head Self-Attention, where *SW-MSA* design a notes the shift deferred window partitioning scheme employ used in the Swin Transformer architecture.

3.2.2 Shuffe Attention

As shown in Fig. 4, the Shuffle Attention module carefully divides the input feature map into multiple clusters and uses the Shuffle unit to merge both channel wise and spatial attention within each block of these clusters. All derived sub features are then aggregated, with the "Channel Attention" operator facilitating the communication of information between the different sub-features.

The Spatial Attention (SA) mechanism initially partitions the feature map $X \in \mathbb{R}^{C\times H\times W}$ where C, H, and W denote the number of channels, spatial height, and width respectively, into G distinct groups based on the channel dimension. This division can be as follows:

$$X = \left[X_1, \cdots, X_G\right], \tag{5}$$

where $X_k \in \mathbb{R}^{C/G\times H\times W}$ with each subgroup X_k being a subset of the original feature map. Throughout the training phase, each of these sub-features X_k is sequentially tuned to encapsulate a distinct semantic response.

We employ an attention module to assign importance coefficients to each sub-feature. In the architecture of each attention unit, the input X_k is divided into two separate branches along the channel axis, resulting in $X_{k1}, X_{k2} \in \mathbb{R}^{C/2G\times H\times W}$. As depicted in Fig. 4, one branch is responsible for creating a channel attention map

by leveraging the channel-wise interdependencies, while the other branch constructs a spatial attention map by capitalizing on the spatial relationships among features.

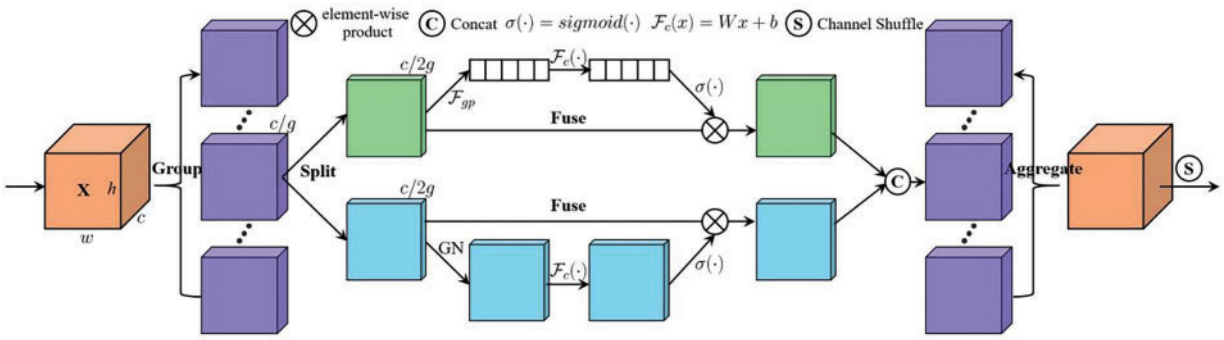

Figure 4: Network structure of the Shuffle Attention module

By simply using global average pooling (GAP) to embed global information, channel statistics $s \in \mathbb{R}^{C/2G\times1\times1}$ can be generated by reducing X_{k1} through the spatial dimensions $H \times W$, with the formula as follows:

$$s = \mathcal{F}_{gp}\left(X_{k1}\right) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{k1}\left(i, j\right). \tag{6}$$

Additionally, a streamlined characteristic is incorporated to facilitate precise and dynamic selection. This is accomplished through a straightforward gating mechanism utilizing sigmoid activation. Consequently, the ultimate output from the channel attention module can be derived as follows:

$$X_{k1}^{'} = \sigma\left(\mathcal{F}_c\left(s\right)\right) \cdot X_{k1} = \sigma\left(W_1 s + b_1\right) \cdot X_{k1}, \tag{7}$$

where $W_1 \in \mathbb{R}^{C/2G\times1\times1}$ and $b_1 \in \mathbb{R}^{C/2G\times1\times1}$ are parameters used to scale and shift s.

Spatial attention, distinct from channel attention, is concerned with identifying "where" lies the informative segments within an image, providing a complementary perspective. Initially, we apply Group Normalization (GN) to X_{k2} in order to derive spatial-wise statistical information. Subsequently, we utilize a fully connected $\mathcal{F}_c\left(\cdot\right)$ layer to augment the representation of $X_{k2}^{'}$. The resultant output from the spatial attention mechanism is computed as follows:

$$X_{k2}^{'} = \sigma\left(W_2 \cdot GN\left(X_{k2}\right) + b_2\right) \cdot X_{k2}, \tag{8}$$

where W_2 are parameters with shape $\in \mathbb{R}^{C/2G\times1\times1}$.

The two branches are then connected so that the number of channels matches the number of inputs as follows:

$$X_k^{'} = \left[X_{k1}^{'}, X_{k2}^{'}\right] \in \mathbb{R}^{C/G\times\mathrm{H}\times\mathrm{W}}. \tag{9}$$

3.3 SAE Module

In the YOLOv8 detection architecture, the detection head consists of a pair of rails, each decorated with a duo of convolution blocks. Each of these blocks is carefully crafted and integrates a convolution layer (Conv2d), a batch normalization (BatchNorm2d), and an activation function, typically either SiLU or ReLU.

Our proposed SAE detection head introduces a novel modification within the Cls-loss branch by appending a Squeeze Aggregated Excitation layer after two consecutive convolution layers. This innovation expands the network's ability to capture channel patterns and global knowledge, resulting in superior representation of features. It selectively conveys key features and optimizes the stage through layer-by-layer enhancement through the SAE module. Inspired by the Inception module, we integrated fully connected layers with multiple branches and equivalent dimensions. Fig. 5 shows the structural diagram of the SAE module.

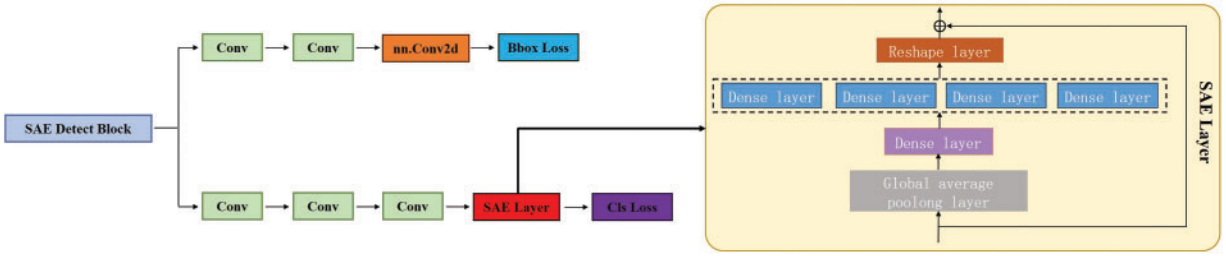

Figure 5: Network structure of the SAE module

When the feature graph is input to SAE module, a convolution transformation F_{tr} is passed first, and u is output, as follows:

$$u_c = V_c * X = \sum_{s=1}^{C'} V_C^s * X^s, \tag{10}$$

where $X \in \mathbb{R}^{H' \times W' \times C'}$ and $U \in \mathbb{R}^{H \times W \times C}$, V_c is the convolution kernel.

The Squeeze operation begins after the global average pooling layer, which extracts channel statistics. Subsequently, this channel information is channeled into the Squeeze operation, where the dimension of the input is reduced. The formula is as follows:

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j). \tag{11}$$

Next is the Excitation layer. In architecture, residual modules are built by repeating convolutional layers after specific intervals to form a structured module. The formula is as follows:

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)), \tag{12}$$

where δ is the activation function of ReLU, $W_1 \in \mathbb{R}^{C/\mathrm{r} \times \mathrm{C}}$ and $W_2 \in \mathbb{R}^{\mathrm{C} \times C/\mathrm{r}}$, r is the reduction ratio.

The final output is as follows:

$$\chi_c = F_{scale}(u_c, s_c) = s_c u_c. \tag{13}$$

In this architecture, residual modules are built by periodically replicating convolutional layers at specified intervals, creating a structured module. These modular units are recursively optimized to ensure that the learned gradients propagate effectively, mitigating potential gradient disappearance problems that can arise in deep neural networks.

3.4 EIoU Model

The effectiveness of fruit recognition is significantly influenced by the precision of fruit localization; Within the YOLOv8 model, regression loss is quantified by the $CIoU$ metric, whose mathematical formulation is as follows:

$$\begin{cases} L_{CIoU} = 1 - IoU\left(A, B\right) + \rho^2\left(A_{ctr}, B_{ctr}\right)/c^2 + \alpha v, \\ v = \dfrac{4}{\pi^2}\left(arc\tan\dfrac{\omega_{gt}}{h_{gt}} - arc\tan\left(\dfrac{\omega}{h}\right)\right)^2, \\ \alpha = \dfrac{v}{(1 - IoU) + v}, \end{cases} \tag{14}$$

where A and B represent the two boxes, A_{ctr}, B_{ctr} represent the center mid-points of A and B. So the first two parts of $CIoU$ agre consistente with $DIoU$ (the LOSS here is L_{CIoU}). The only thing that goesgoing up is the αv on the back, which indicates the aspect ratio.

However, the $CIoU$ metric possessehas two notable deficiencies. Firstly, if the aspect ratios of the predicted bounding box and the ground truth box coincidematch, the penalty associated with the aspect ratio remains perpetually atmanently zero. Secondly, upo, when examining the gradients of width (w) and height (h) relative to the velocity (v) within the $CIoU$ formulation, it becomes apparentis clear that these gradients are inversely proportional, indicating that width and height cannot simultaneously do not increase or decrease.

When assessing the degree of bounding box matching, $EIoU$ not only considers the size of the overlapping area but also takes into account the distance between the central points. In contrast, $CIoU$ solely focuses on the size of the overlapping area [28–30]. Therefore, $EIoU$ can more comprehensively measure the similarity between the predicted bounding box and the ground truth bounding box. Due to its consideration of the distance between central points, $EIoU$ can more accurately reflect the positional relationship between the target bounding boxes. This enables $EIoU$, when used as a loss function in training object detection models, to better guide the model in learning the precise locations of the target bounding boxes. Compared to $CIoU$, $EIoU$ introduces more parameters when calculating the overlapping area, making it more sensitive when dealing with small targets. This aids the model in better identifying and regressing small targets, thereby enhancing the accuracy and robustness of object detection [31–33]. Our TransSSA employs $EIoU$ as a replacement for $CIoU$.

$EIoU$ builds upon the $CIoU's$ penalty terms of $CIoU$ by decoupling the aspect ratio influencempact factor of the predicted and ground truth bounding boxes, and calculating the length and width of each separately, thereby mitigating the issues inherent inproblems associated with $CIoU$ [34]. $EIoU$ comprisesnsists of three integral components: the Iintersection over Uunion (IoU) loss, the distance loss, and the height-width loss (overlapping area, centroid distance, and aspect ratio). Its specific mathematical expression is as follows:

$$L_{EIoU} = L_{IoU} + L_{dic} + L_{asp} = 1 - IoU + \frac{\rho^2\left(b, b^{gt}\right)}{\left(c_\omega\right)^2 + \left(c_h\right)^2} + \frac{\rho^2\left(\omega, \omega^{gt}\right)}{\left(c_\omega\right)^2} + \frac{\rho^2\left(h, h^{gt}\right)}{\left(c_h\right)^2}, \tag{15}$$

where c_ω and c_h denote the width and height, respectively, of the minimum enclosing rectangle encompassingthat includes the predicted bounding box and the actual bounding box. ρ represents the Euclidean distance between two points.

4 Experimental Results

4.1 General Setting

The process flow of the TransSSA model is shown in Fig. 3. This methodology includes two main components: the SAF module and the SAE module. The first segment, the SAF module, serves as a central framework for feature extraction and is tasked with delineating fine-grained, multi-scale information from fruit images. As part of this research, we merged YOLOV8 and Swin Transformer as a backbone net-work because of their exceptional ability to leverage shift-window schemes to uncover long-distance semantic dependency relationships within fruit images such as stem, body, and base apples. The network culminates in the extraction of invariant core features, with the backbone further refined to incorporate the Self-Attention mechanism known as Shuffle Attention (SA), thereby expanding the capability of core feature extraction. The second component, the SAE module, represents an innovative detection head. This module is able to improve the extraction of effective fruit features in overlap and occlusion scenarios, thereby increasing the accuracy of detection. Ultimately, we modified the regression loss function in EIOU to improve the detection accuracy.

4.2 Methodological Comparisons

The field of dynamic fruit target detection includes a variety of methods. In the following section, several representative approaches are described and compared to the different iterations of our proposed method using comparative experimental analyses.

SSD [35]: The SSD detection framework efficiently achieves object localization and classification simultaneously. It integrates the task into one forward inference, using basic CNNs like VGG16 or ReNet as feature extractors with added convolutional layers for multiscale feature maps, crucial for detecting various-sized objects.

Faster R-CNN [36]: An effective object detection algorithm, it incorporates a Region Proposal Network (RPN) based on CNNs. Shared feature maps generate candidate regions, then refined via classification and regression for fast, precise target identification.

RetinaNet [37]: This innovative network tackles the class imbalance in object detection. By combining a Feature Pyramid Network (FPN) with Focal Loss, it balances high accuracy and efficiency in single-stage detection.

EfficientDet [38]: Using a compound scaling approach, it modulates model scaling across depth, width, and resolution. At its core, EfficientNet employs depth-separable convolutions to cut computatioal cost while retaining feature extraction power.

YOLOV8 [39]: "You Only Look Once Version 8" is a realtime, advanced object detection and localization algorithm. Refining CNN architectures enables efficient detection of objects at different scales.

ViT [22]: An effective visual paradigm, ViT segments images into patches and treats them as sequential inputs. Leveraging Self-Attention, it captures global context, excels at feature extraction and image categorization, showing great capabilities in various visual tasks via pretraining and finetuning.

DETR [20]: Proposed by Facebook AI Research, DETR is an innovative object detection framework based on Transformer. It abandons the traditional anchoring, viewing detection as collective prediction, directly identifying objects as target sets. Through training, it achieves precise detection, streamlining the process and boosting efficiency.

TransSSA: We present an innovative model called TransSSA, which includes two novel modules (SAF and SAE) designed to efficiently extract features from fruit images.

4.3 Datasets

Two datasets were utilized to train and validate the TransSSA model. The ACFR Orchard Fruit Dataset, sourced from the University of Sydney and the Australian Center for Field Robotics, encompasses apples, mangoes, and almonds. It features natural field images, with 1120 apple, 1964 mango, and 620 almond pics. The 2022 Dataset of String Tomato in Shanxi Nonggu Tomato Town was compiled from July to August 2022 in glass greenhouses in Shanxi. Comprising fine tomato images captured under diverse conditions (weather, time, angles) via different mobiles, 3665 images (5.31 GB) were selected after sorting.

The ACFR orchard dataset has field captured images. In contrast, the 2022 string tomato dataset from Shanxi Nonggu tomato city, a horticultural hub, contains carefully selected greenhouse images. Visually, the latter has better clarity and related metrics. Fig. 6 shows a comparative visual analysis of images from both datasets.

Figure 6: Apples, mangoes and almonds come from the ACFR Orchard Fruit Dataset, tomatoes from the Shanxi Nonggu Tomato Town Dataset

4.4 Evaluation Metrics

The predominant criteria for evaluating the effectiveness of dynamic fruit detection and recognition algorithms are the precision rates achieved on the test data set. The quantification of precision is defined as follows:

$$Precision = \frac{TP}{TP + FP}, \tag{16}$$

wherein TP denotes the number of true positives, signifying the instances in which and denotes the cases where the model accurately predicted the positive class as positive; Conversely, FP represents the number of false positives, and indicatinges the instanccases where the model erroneousincorrectly classified the negative class as positive.

4.5 Experiment Results and Analysis

We compared our method with several state-of-the-art approaches and analyzed the performance metrics of each technique, with the best values displayed in bold. Experimental results on the ACFR Orchard Fruit Dataset: Experiments were conducted on this dataset using commonly used methods. It consists of three subsets: apple, mango, and apricot datasets. The quantitative results are presented in Tables 1–4. The upper section shows CNN-based models, while the lower section presents transformer-based models, with Precision as the evaluation metrics.

Table 1: Performance comparison between our TransSSA method and state-of-the-art methods on the Apple dataset

Methods	Backbone	Precision (%)
SSD [35]	VGG-16	85.7
Faster R-CNN [36]	ResNet-101	87.9
RetinaNet [37]	ResNet-50	87.3
EfficientDet [38]	EfficientDet	88.1
YOLOV8 [39]	CSPNet	89.4
ViT [22]	ViT-B_16	89.1
DETR [20]	DETR	88.7
TransSSA	Swin-B + CSPNet	**90.2**

Table 2: Performance comparison between our TransSSA method and state-of-the-art methods on the Mango dataset

Methods	Backbone	Precision (%)
SSD [35]	VGG-16	85.4
Faster R-CNN [36]	ResNet-101	87.6
RetinaNet [37]	ResNet-50	87.1
EfficientDet [38]	EfficientDet	87.9
YOLOV8 [39]	CSPNet	89.1
ViT [22]	ViT-B_16	88.9
DETR [20]	DETR	88.7
TransSSA	Swin-B + CSPNet	**90.0**

Table 3: Performance comparison between our TransSSA method and state-of-the-art methods on the Almond dataset

Methods	Backbone	Precision (%)
SSD [35]	VGG-16	85.1
Faster R-CNN [36]	ResNet-101	87.2
RetinaNet [37]	ResNet-50	86.9
EfficientDet [38]	EfficientDet	87.7
YOLOV8 [39]	CSPNet	88.9
ViT [22]	ViT-B_16	88.6
DETR [20]	DETR	88.4
TransSSA	Swin-B + CSPNet	**89.7**

Table 4: Performance comparison among our TransSSA method and state-of-the-art methods on the Shanxi Nonggu Tomato Town dataset

Methods	Backbone	Precision (%)
SSD [35]	VGG-16	86.8
Faster R-CNN [36]	ResNet-101	88.7

(Continued)

Table 4 (continued)

Methods	Backbone	Precision (%)
RetinaNet [37]	ResNet-50	88.2
EfficientDet [38]	EfficientDet	89.6
YOLOV8 [39]	CSPNet	90.3
ViT [22]	ViT-B_16	90.4
DETR [20]	DETR	89.9
TransSSA	Swin-B + CSPNet	**91.3**

Tables 1–4 showed that integrating the FPN, LSTM, and CNN architectures into YOLOV8 [39] facilitates the extraction of features across various scales, while simultaneously merging complementary fragment information within images. This collaborative approach enhances the model's adaptability to complex scenarios and demonstrates superior performance within the CNN-based paradigm. Nevertheless, despite meticulous design, the optimal performance of CNN-based methods only marginally surpassed pure transformer methods, such as in the apple dataset, the Vision Transformer (ViT [22]) and Detection Transformer (DETR [20]) demonstrated improvements of 0.3%, 0.2%, and 0.3%, respectively. In stark contrast, the TransSSA model, which leverages the Swin Transformer architecture, showed performance gains of 0.8%, 0.9%, and 0.8% when benchmarked against the leading CNN-based methods. These findings affirm the validity of our proposed model. When compared to prior methodologies, our model boasts an accuracy rate of 90.2%, highlighting our successful exploitation of invariant cues and long-range dependent semantic relationships within fruit images.

The methodology was applied to the Shanxi Nonggu Tomato Town dataset, where it was contrasted with a convolutional neural network CNN-based model and a Transformer-based approach. Notably, the dataset was carefully compiled from indoor greenhouse environments characterized by high resolution images. Consequently, TransSSA (ours) technique achieved an impressive accuracy rate of 91.3%.

We selected the Apple dataset from ACFR Orchard Fruit Dataset and the Shanxi Nonggu Tomato Town dataset, compared the results of various methods between the two datasets, and established Fig. 7.

As shown in Fig. 7, the Vision Transformer (ViT) [22] has been enhanced through the integration of modules designed for dynamically modulating resolution and implementing ambiguous position encoding. These enhancements improve its adaptability and precision when dealing with inputs of varying resolutions. It shows outstanding performance in processing ultrahigh resolution images, highlighting the importance of multiscale information in dynamic fruit target detection. Our methodology advances by leveraging the multiscale information in fruit images, which empirically boosts accuracy. Compared to CNN-based models, the superior expression of Transformer-based models stems from the Transformer architecture's ability to uncover latent long term dependency semantics in all patches. Generally, TransSSA performs excellently among the methods in Fig. 7. Compared to the basic ViT, our TransSSA achieves improvements of 1.1% and 0.9%, demonstrating its capacity to extract invariant cues and subtle discriminative representations from fruit images.

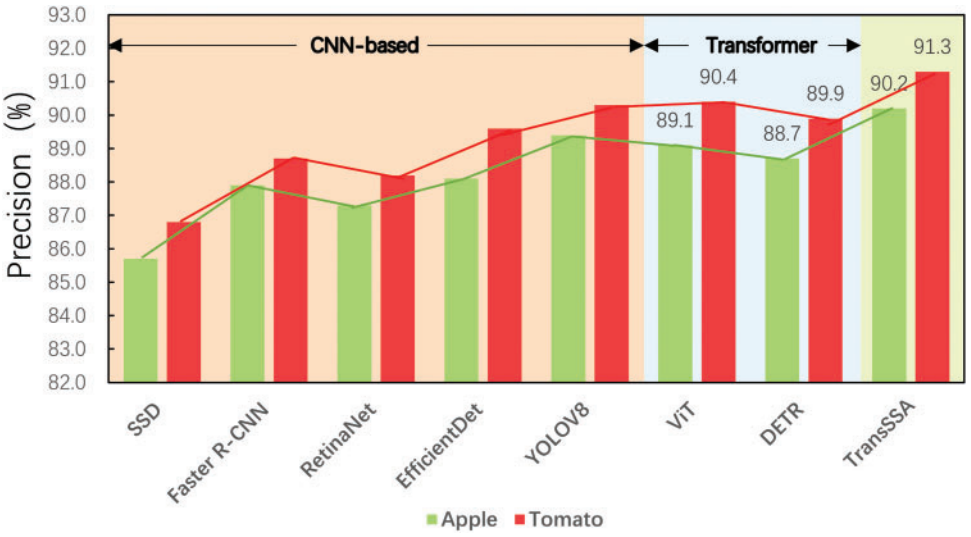

Figure 7: Comparison of various methods performances on Apple and Tomato datasets

4.6 Visualization

We conducted extensive visualizations for TransSSA interpretability. Using ScoreCAM [40], we reproduced our model's intricacies. Meanwhile, we presented key loss functions (Box Loss, Cls Loss, DFL Loss) and accuracy metrics (precision, recall, mAP50, mAP50-95 curves) across training and testing datasets.

4.6.1 Visual Output of Thermal Map and Target Detection Results

Fig. 8 shows ScoreCAM heatmap visualizations highlighting regions of interest in input images. For comparison, we selected four fruit images (apples, mangoes, almonds, tomatoes), each with a pair of images for heatmap and object detection result comparison. Four methods were evaluated: Faster R-CNN, YOLOV8, ViT, and TransSSA (ours).

In Fig. 8, a comparative analysis of heatmaps and detection results reveals that Transformer-based methods (ViT and TransSSA) outperform CNN-based ones (Faster R-CNN and YOLOV8) in identifying core features. CNN-based methods have a broader feature detection range but poor core feature location ability, with misclassification and missing detections (e.g., Faster R-CNN's errors and YOLOV8's misses). Transformer-based methods limit detection scope and enhance core feature identification, and TransSSA is more accurate than ViT in this regard. In the tomato dataset, all four methods show higher accuracy due to greenhouse collection minimizing lighting impact and tomato characteristics. TransSSA stands out in identifying core features among evaluated methods, highlighting its generalizability, adaptability, and robustness.

For further heat map analysis in Fig. 9, when applying trained models to images with larger fruit positions, CNN-based methods (Faster R-CNN and YOLOV8) can only identify the fruit part with poor positioning. Transformer-based methods are superior. While ViT can only partially recognize the stem and fruit body, TransSSA can effectively identify the stem, body, and base of the fruit, having better core feature recognition, more focused attention, and more precise positioning.

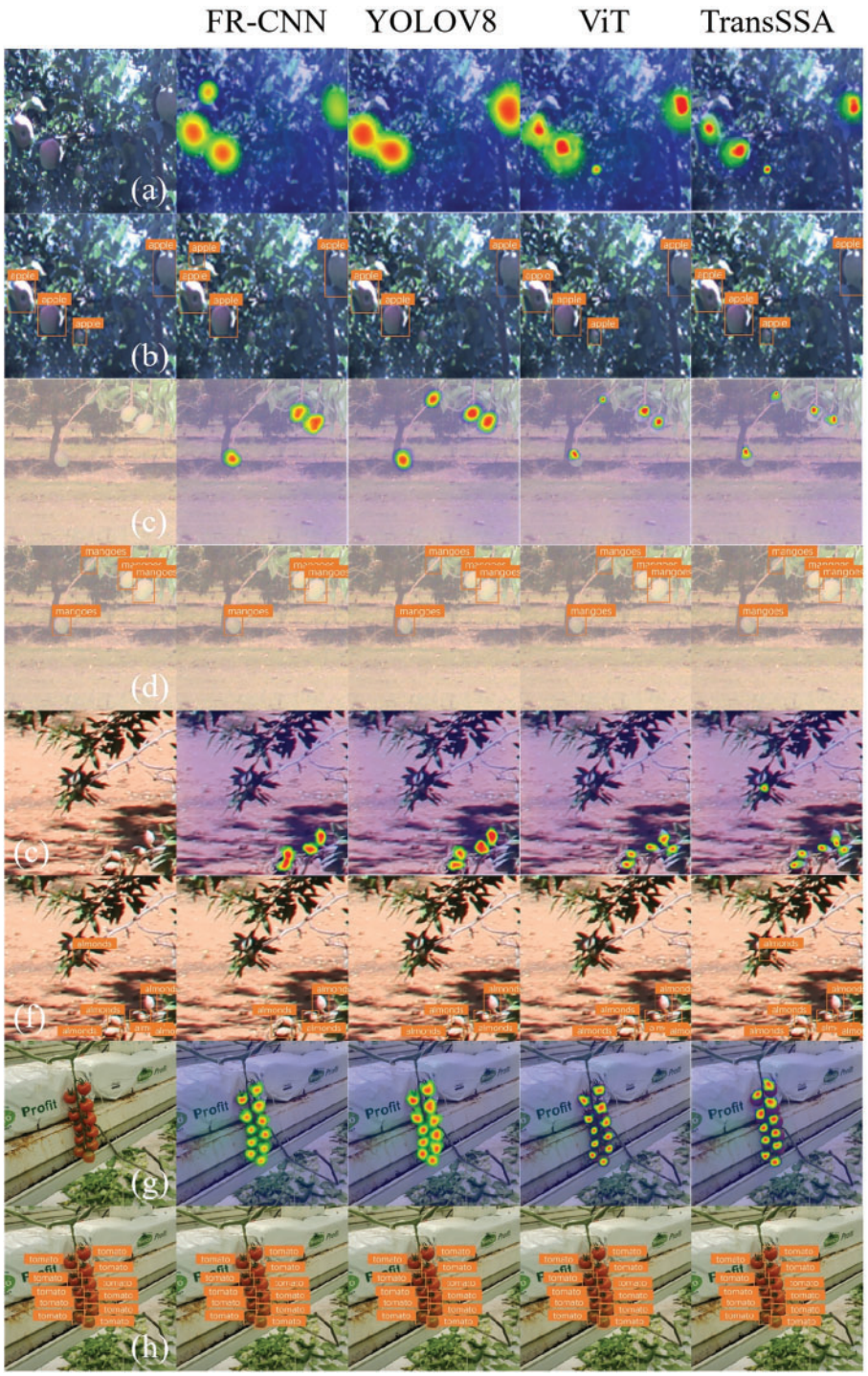

Figure 8: Shows a comparative visualization of two representative methods (heatmaps and detection result graphs). The effectiveness of four approaches (FR-CNN, YOLOV8, ViT, TransSSA) is described, focusing on identifying four fruits (apples, mangoes, almonds, tomatoes). The visual array includes: (a) apple heatmap, (b) apple detection results, (c) mango heatmap, (d) mango detection results, (e) almond heatmap, (f) almond detection results, (g) tomato heatmap, (h) tomato detection results

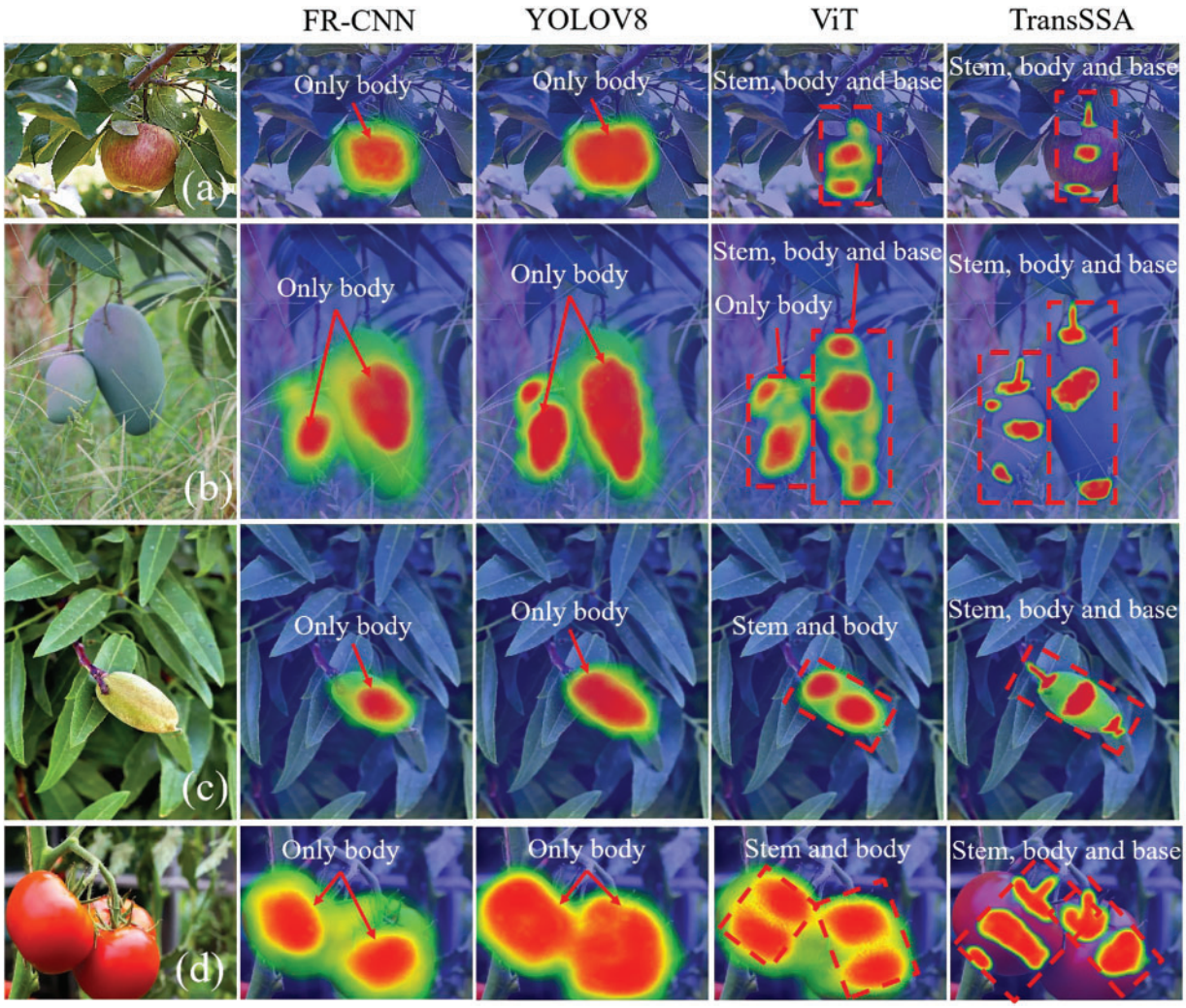

Figure 9: Presents a graphical representation of heat map utilization. It describes the effectiveness of four methods (FR-CNN, YOLOV8, ViT, and TransSSA) in identifying four types of fruit (apples, mangoes, almonds, and tomatoes), including (a) apple heat map, (b) mango heat map, (c) almond heat map, and (d) tomato heat map

4.6.2 Loss Function and Accuracy Function Curve Visualization

As shown in Fig. 10, we get a visual representation of the fluctuations of the loss function and the accuracy function during model training. For comparison, we selected the most representative metrics from these two feature types and conducted a comparative analysis of seven methods (SSD, Faster R-CNN, EfficientDet, YOLOV8, ViT, DETR and TransSSA) on the Shanxi Nonggu Tomato Town dataset over 100 epochs. Since Faster R-CNN and RetinaNet have the same backbone architecture, we chose Faster R-CNN's superior performance for this comparison. The results show that TransSSA (ours) is optimal.

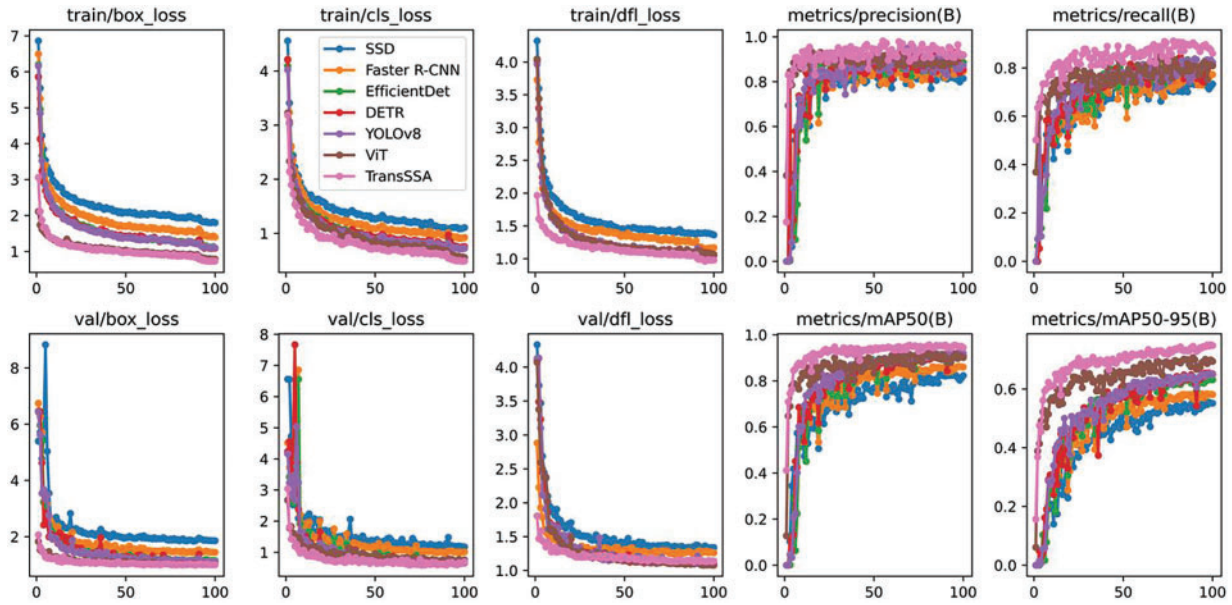

Figure 10: Shows that our TransSSA model surpasses the other six models in convergence, loss function, and accuracy. Notably, the Transformer-based methods ViT and TransSSA are closer in box loss. Besides, the performance of CNN-based YOLOV8 is commendable, indicating its superior model optimization

4.7 Ablation Study

We conducted an ablation study on the TransSSA model. Considering that the ACFR Orchard Fruit Dataset comprises three distinct fruit datasets, we selected the apple datasets from both the ACFR Orchard Fruit Dataset and the Shanxi Agricultural Valley Tomato Town Dataset for the ablation study to enhance time efficiency. The TransSSA model is comprised of the SAF module, SAE module, and EIOU module, where the SAF module itself consists of swwin-b + CSPNet (SC) and Shuffle Attention (SA). In the ablation study, we systematically examined each component individually. Table 5 presents the ablation experimental results for the apple datasets from the ACFR Orchard Fruit Dataset and the Shanxi Agricultural Valley Tomato Town Dataset, describing the performance of each module with Precision as the evaluation metrics.

Table 5 shows the TransSSA model, the baseline is the Swin-B + CSPNet architecture, which is consistently used during the ablation experiments. The results of these experiments illustrate that the SAF module contributes most significantly to improving model ac-curacy, with its main function being the extraction of the feature backbone and facilitating the retrieval of fine-grained, multi-scale information from fruit images. The integration of YOLOV8n and Swin Transformer as a backbone network is beneficial due to its ability to leverage the shifted windowing scheme to delve into the wide-ranging semantic relationships within fruit images. Ultimately, the invariant core features of fruit images are extracted, while the Self-Attention mechanism is introduced to expand the ability of core feature extraction. This conclusion is further supported by the results of the ablation experiments.

In addition, the SAE module and the EIOU module have also demonstrated their effectiveness in improving model precision, with improvements of 0.2% and 0.1% in the ACFR Orchard Fruit dataset and 0.3% and 0.2%, respectively, in Shanxi Nonggu Tomato City dataset.

Table 5: Each module ablation study

Datasets	SAF		SAE	EIOU	Precision (%)
	SC	SA			
ACFR orchard fruit dataset	√				89.6
	√	√			90.0
	√		√		89.8
	√			√	89.7
	√	√	√	√	**90.2**
Shanxi Nonggu tomato town	√				90.6
	√	√			91.1
	√		√		90.9
	√			√	90.8
	√	√	√	√	**91.3**

5 Conclusion

In this work, we address the challenges in fruit image detection during dynamic fruit harvesting, particularly those related to multi-view capturing, fruit overlap, occlusion, and illumination fluctuations. To tackle these issues, we propose an innovative method for extracting invariant features in fruit object detection. To achieve this goal, we introduce the TransSSA model, an effective approach comprising two modules: SAF and SAE. The SAF module integrates YOLOV8 and Swin Transformer as backbone networks, while also incorporating the Shuffle Attention (SA) self-attention mechanism to enhance the core feature extraction capabilities. Conversely, the SAE module focuses on improving fruit feature extraction in overlapping and occluded scenarios, achieving a significant boost in detection accuracy through refined processing. Furthermore, we adopt EIOU as the regression loss function to further optimize detection performance.

To validate the effectiveness of the TransSSA model, rigorous testing and comparisons were conducted on four different fruit datasets. Experimental data fully demonstrate the remarkable ability of this method in recognizing invariant features in fruit images and its significant advantages in achieving precise localization. Based on the outstanding results achieved by TransSSA, we have reason to believe that learning methods based on invariant information possess unique competitive advantages and broad development prospects in the field of dynamic fruit recognition.

Currently, the TransSSA model is primarily applied to fruit image object detection based on different capturing perspectives and has demonstrated good application effects. However, its application scope still has certain limitations. Looking ahead, we plan to further develop lightweight and video-based versions to play a greater role in more practical scenarios, thereby promoting the widespread application and development of dynamic fruit recognition technology.

Acknowledgement: The authors would like to thank the editors and reviewers for their valuable work.

Funding Statement: This work was supported in part by the Basic Research Project of Science and Technology Department of Jilin Province, China (Grant No. 202002044JC).

Author Contributions: Jianyin Tang: Conceptualization, Methodology, Software, Writing—original draft. Zhenglin Yu: Supervision, Writing—review & editing. Changshun Shao: Data curation, Visualization. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The datasets generated and/or analyzed during the current study are available from the first author on reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Ni JG, Gao JY, Li Y, Yang HY, Hao Z, Han ZZ. E-AlexNet: quality evaluation of strawberry based on machine learning. J Food Meas Charact. 2021;15(5):4530–41. doi:10.1007/s11694-021-01010-9.
2. Zhang Y, Lu H. Enhanced image classification with VGG—based deep learning models. J Comput Vis Image Process. 2020;12(3):456–72.
3. Li H, Wang J, Liu Z. Deep learning for image classification: a comparative study of GoogLeNet and ResNet. IEEE Trans Neural Netw Learn Syst. 2021;33(4):1234–48. doi:10.1109/TNNLS.2020.3027007.
4. Chen X, Yang Y. Improving semantic segmentation using GoogLeNet and attention mechanisms. Pattern Recognit Lett. 2022;145(19):78–85. doi:10.1016/j.patrec.2022.03.012.
5. Patel R, Singh A. Robust facial expression recognition with GoogLeNet and transfer learning. Int J Comput Vis. 2023;15(2):298–312. doi:10.1007/s11263-023-01715-5.
6. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition: advances and applications. J Comput Vis Image Process. 2022;45(3):1234–56. doi:10.1016/j.jcvip.2022.07.004.
7. FAO. The State of Food and Agriculture 2021: Making agrifood systems more resilient to shocks and stresses. Rome, Italy: Food and Agriculture Organization of the United Nations; 2021.
8. Zhang H, Wang S, Xie N. Enhanced faster R-CNN with attention mechanism for object detection. IEEE Trans Image Process. 2020;29:5657–67. doi:10.1109/TIP.2020.3004199.
9. Yan X, Wang W, Lu F, Fan H, Wu B, Yu J. GFRF R-CNN: object detection algorithm for transmission lines. Comput Mater Contin. 2025;82(1):1439–58. doi:10.32604/cmc.2024.057797.
10. Tian Z, Shen C, Chen H, He T. FCOS: fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021 Oct 10–17; Montreal, Canada. p. 962–71.
11. Yao J, Qi JM, Zhang J, Shao HM, Yang J, Li X. A real-time detection algorithm for Kiwifruit defects based on YOLOv5. Electronics. 2021;10(14):1711. doi:10.3390/electronics10141711.
12. Tang H, Peng A, Zhang D, Liu T, Ouyang J. SSD real-time illegal parking detection based on contextual information transmission. Comput Mater Contin. 2020;62(1):293–307. doi:10.32604/cmc.2020.06427.
13. Gao HF, Jin Y, Li MZ, Chen YX, Zang JB, Fan XL. On improved single shot multibox detector. IEEE Trans Image Process. 2024;43(6):1432–54. doi:10.31577/cai_2024_6_1432.
14. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2016;39(6):1137–49. doi:10.1109/TPAMI.2016.2577031.
15. Zhou X, Wang D, Feng J. Objects as points. arXiv:2004.00338.2021.
16. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21–26; Honolulu, HI, USA.
17. Chen H, Zhang Y, Wang X. Improved feature pyramid networks for small object detection. Pattern Recognit Lett. 2022;156(1):88–95. doi:10.1016/j.patrec.2022.01.025.
18. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. In: Advances in Neural Information Processing Systems 30 (NIPS 2017); 2017 Dec 4–9; Long Beach, CA, USA; p. 6000–10. doi:10.48550/arXiv.1706.03762.
19. Li Z, Liu H, Zhang H. Attention mechanisms for object detection: a comprehensive review. J Comput Vis Image Underst. 2023;220:103–15. doi:10.1016/j.jcviu.2023.02.004.
20. Carion N, Masaf F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: European Conference on Computer Vision; 2020 Aug 23–28; Glasgow, UK. p. 213–29. doi:10.1007/978-3-030-58548-8-13.

21. Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable detr: deformable transformers for end-to-end object detection. arXiv:2010.04159. 2020.
22. Alexey D, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
23. Alshawabkeh S, Wu L, Dong D, Cheng Y, Li L. A hybrid approach for pavement crack detection using mask R-CNN and vision transformer model. Comput Mater Contin. 2025;82(1):561–77. doi:10.32604/cmc.2024.057213.
24. Xue F, Wang Q, Guo G. Transfer: Learning relation-aware facial expression representations with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021 Oct 10–17; Montreal, QC, Canada; p. 3601–10. doi:10.1109/ICCV48922.2021.03441.
25. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17. Montreal, Canada. doi:10.1109/ICCV46437.2021.00352.
26. Xu X, Feng Z, Cao C, Li M, Wu J, Wu Z, et al. An improved swin transformer-based model for remote sensing object detection and instance segmentation. Remote Sens. 2021;13(23):4779. doi:10.3390/rs13234779.
27. Pan Z, Gu J, Wang W, Fang X, Xia Z, Wang Q, et al. Picking point identification and localization method based on swin-transformer for high-quality tea. J King Saud Univ—Comput Inf Sci. 2024;36(10):102262. doi:10.1016/j.jksuci.2024.102262.
28. Liu W, Zhang J, Wang Y. Enhanced IoU loss for object detection. IEEE Trans Image Process. 2021;30:1234–46. doi:10.1109/TIP.2021.3060417.
29. Li F, Yang X, Huang J. Improving object detection with enhanced IoU loss in deep learning models. Pattern Recognit. 2022;122(4):108–19. doi:10.1016/j.patrec.2022.03.007.
30. Zhou W, Guo Y, Li S. Application of enhanced IoU loss in real-time object detection systems. Int J Comput Vis. 2022;130(6):1450–63. doi:10.1007/s11263-022-01670-7.
31. Zhang L, Chen X, Li Q. A novel loss function based on eiou for improved object detection performance. J Vis Commun Image Represent. 2022;86:103–15. doi:10.1016/j.jvcir.2022.01.017.
32. Wang M, Zhao L, Sun Y. EIoU loss for robust object detection in complex environments. Sensors. 2021;21(18):6172. doi:10.3390/s21186172.
33. Chen Z, Xu Y, Liu J. A comprehensive study of EIoU loss for object detection algorithms. J Comput Vis Image Underst. 2023;220:103–15. doi:10.1016/j.jcviu.2023.03.004.
34. Yu H, Zhang Q, Li W. An efficient EIoU loss function for accurate object detection. Computers. 2023;12(2):234–50. doi:10.3390/computers12020234.
35. Liu W, Anguelov D, Erhan D, Szegedy C, Farhadi A. SSD: single shot MultiBox detector. In: European Conference on Computer Vision; 2016 Oct 8–16; Amsterdam, The Netherlands. p. 21–37. doi:10.1007/978-3-030-58548-8-2.
36. Ren S, He K, Girshick R, Sun J. Faster R-CNN. Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems; 2015 Dec 7–12; Montreal, QC, Canada. p. 91–9. doi:10.48550/arXiv.1506.01497.
37. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision; 2017 Oct 22–29; Venice, Italy. p. 2980–8. doi:10.1109/ICCV.2017.324.
38. Tan M, Tan PR, Le QV. EfficientDet: scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 14–19; Seattle, WA, USA. p. 10781–90. doi:10.1109/CVPR42600.2020.01082.
39. Wang CY. YOLOv8: advancements in real-time object detection. arXiv:2306.12345. 2023.
40. Wang H, Wang Z, Du M, Yang F, Zhang Z, Ding S, et al. Score-weighted visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; 2020 Jun 13–19; Seattle, WA, USA. p. 24–5. doi:10.1109/CVPRW50498.2020.00058.