*Journal of Autonomous Intelligence (2024) Volume 7 Issue 3*
doi: 10.32629/jai.v7i3.1189

ORIGINAL RESEARCH ARTICLE

# CETR: CenterNet-Vision transformer model for wheat head detection

**K. G. Suma[1], Gurram Sunitha[2,*], Ramesh Karnati[3], E. R. Aruna[3], Kachi Anvesh[4], Navnath Kale[5], P. Krishna Kishore[6]**

[1] *School of CSE, VIT-AP University, Amaravati, Andhra Pradesh 522237, India*
[2] *Department of AI & ML, School of Computing, Mohan Babu University, Tirupati 517102A.P., India*
[3] *Department of CSE, Vardhaman College of Engineering, Hyderabad 501218, India*
[4] *Department of IT, Vardhaman College of Engineering, Hyderabad 501218, India*
[5] *School of Computer Engineering, MIT Academy of Engineering, Alandi 411052, India*
[6] *Department of IT, BVRIT Hyderabad College of Engineering for Women, Hyderabad 500090, India*
*** Corresponding author:** Gurram Sunitha, gurramsunitha@gmail.com

## ABSTRACT

Wheat head detection is a critical task in precision agriculture for estimating crop yield and optimizing agricultural practices. Conventional object detection architectures often struggle with detecting densely packed and overlapping wheat heads in complex agricultural field images. To address this challenge, a novel CEnternet-vision TRansformer model for Wheat Head Detection (CETR) is proposed. CETR model combines the strengths of two cutting-edge technologies—CenterNet and Vision Transformer. A dataset of agricultural farm images labeled with precise wheat head annotations is used to train and evaluate the CETR model. Comprehensive experiments were conducted to compare CETR's performance against convolutional neural network model commonly used in agricultural applications. The higher mAP value of 0.8318 for CETR compared against AlexNet, VGG19, ResNet152 and MobileNet indicates that the CETR model is more effective in detecting wheat heads in agricultural images. It achieves a higher precision in predicting bounding boxes that align well with the ground truth, resulting in more accurate and reliable wheat head detection. The higher performance of CETR can be attributed to the combination of CenterNet and ViT as a two-stage architecture taking advantage of both methods. Moreover, the transformer-based architecture of CETR enables better generalization across different agricultural environments, making it a suitable solution for automated agricultural applications.

***Keywords:*** wheat head detection; CenterNet; vision transformer; object detection; smart agriculture

**ARTICLE INFO**

Received: 28 August 2023
Accepted: 4 December 2023
Available online: 2 January 2024



## 1. Introduction

Wheat is indeed a heavily explored crop as it is a significantly important staple food source worldwide. Wheat head detection from farm images is a critical task in precision agriculture and crop monitoring[1]. It involves identifying and localizing individual wheat heads within agricultural field images. Accurate wheat head detection is essential for various applications, such as estimating crop yield, assessing crop health, optimizing irrigation, and implementing targeted interventions. Advanced computer vision techniques are significantly used for wheat head detection from farm images. The most popular object detection techniques are deep learning-based models.

Deep neural networks are now considered to be the systems of artificial intelligence in use today[2,3]. Various tasks have traditionally required different types of neural networks. Traditional deep learning models are generally sufficient for handling a wide range of data types

and tasks. They can handle image data, text data, tabular data, multimedia data, time series data, sensor data etc. Well-designed and fine-tuned traditional deep learning models have been highly successful in many application domains. They took leverage of knowledge learned from large-scale datasets and adapted it to new tasks with limited labeled data. However, when dealing with shift-invariant data, such as image data, the deep model's behavior shall not depend on the absolute position of features in the image but on their relative positions. In shift-invariant deep learning models, the algorithm shall be capable of detecting objects irrespective of their location within the image. This property is crucial for various computer vision tasks because objects of interest may appear at different locations in different images. Shift-invariance allows the model to recognize the same object regardless of where it is situated within the input image.

Many different approaches for solving object detection problems have been developed by using different deep learning methodologies[2,4]. The objective of this research paper is to develop a two-stage deep model to locate wheat heads quickly and precisely from agricultural farm images. This research proposes an efficient two-stage object detection model CEnternet-vision TRansformer (CETR) for wheat head detection from agricultural farm images. The first stage of CETR uses CenterNet deep architecture[3] and the second stage uses a Vision Transformer (ViT)[5]. In the proposed model, object detection is viewed as a problem with explicit object detection. By efficiently eliminating the need for multiple manually constructed components, such as a non-maximal suppression mechanism or anchor creation of the task, the proposed approach optimizes the object detection pipeline. CETR is a methodology that relies on a transformer encoder-decoder structure with an CenterNet objectness loss and CenterNet Center Regression Loss that drives accurate predictions. CETR simultaneously generates the final number of predictions by inferring associations between the objects and the visual context in the wheat farm images. Unlike traditional object detection methods that use region proposal networks, CETR directly predicts object bounding boxes and class labels from a set of fixed-length object detection queries.

The aim of this research is to show how accurate are the vision transformer based deep learning models for object detection in comparison with traditional deep learning methods. The proposed model eliminates the need for manual anchor box design and non-maximum suppression, making the framework highly efficient and accurate. Utilizing a CenterNet objectness loss and CenterNet Center Regression Loss, the model is trained from beginning to end applying a combination of classification and regression losses. When compared to traditional convolutional neural networks, CETR demonstrates its potential as a powerful and robust object detection framework wheat head detection from agricultural farm images. This suggests that CETR can offer significant advantages in terms of accuracy and efficiency for agricultural applications, ultimately contributing to improved crop management and productivity.

## 2. Related work

Deep learning has become a revolutionary technology with wide-ranging applications and significant importance across various domains. Some of the key fields for the importance of deep learning are speech recognition, healthcare, agriculture, finance, economy etc.[6–12].

In 2020, Facebook's research team made significant advancements in computer vision research. They introduced a novel approach to object detection using transformer architecture. Transformers are originally popular in natural language processing[13]. The proposed model presented an end-to-end detection pipeline. This pipeline seamlessly integrates the entire process of object detection, panoptic segmentation, and object identification. This development marked the first-time transformers were effectively incorporated as a fundamental component in the object detection pipeline.

Deep learning plays a critical role in precision agriculture by processing data from various sources, including drones, satellites, and IoT devices. This data is used for tasks like soil analysis, irrigation control,

and crop monitoring. Deep learning models, including CNNs and RNNs, have been employed to detect diseases in crops based on leaf images. These models can accurately identify diseases such as rust, blight, and wilt, enabling timely intervention and reduced crop loss[14]. Deep learning techniques have been applied to identify and manage weeds in agricultural fields. They have been used for accurate weed classification, which enables targeted herbicide application. Crop phenotyping extracts features from plant images and analyzes growth patterns, stress responses, and other traits. This assists breeders in developing more resilient and high-yielding crop varieties. Deep learning is used to build models that simulate complex agroecological processes. These include crop growth, soil nutrient dynamics, and climate interactions. These models assist in sustainable agricultural practices.

Vision transformers have shown great promise in various computer vision tasks such as medical imaging, agriculture imaging etc.[15]. The ability of vision transformers to process large-scale agricultural imagery, capture global contextual information, and handle complex spatial relationships makes them well-suited for various tasks in agricultural imaging. Vision transformers are anticipated to play a bigger part as technology develops in revolutionising sustainable agricultural methods and precision agriculture[16–18].

CenterNet is based on a simple architecture that directly predicts object centers and their corresponding bounding box offsets[3]. This simplicity makes it computationally efficient compared to more complex object detection models, which is beneficial for processing large-scale agricultural images. In order to predict accurate bounding boxes for precise wheat head identification, accurate localisation of each wheat head's centre is essential. CenterNet is specifically designed to predict accurate center points of objects. CenterNet's ability to predict center points makes it more robust to object occlusion, allowing it to handle overlapping wheat heads effectively. CenterNet's attention mechanism enables it to focus on relevant object centers, filtering out clutter and improving the quality of feature maps. CenterNet can be easily scaled to handle various image resolutions and aspect ratios. This flexibility is advantageous when dealing with different types of agricultural farm images with varying sizes and orientations. CenterNet captures global contextual information from the entire image, which is essential for understanding the spatial relationships between wheat heads. This contextual awareness can improve the overall performance of wheat head detection. CenterNet allows for end-to-end training, where the entire model is trained jointly for both object center prediction and bounding box regression. This joint training simplifies the training process and often leads to better performance.

Agricultural farm images can be cluttered and contain various objects and vegetation. Vision transformer's ability to attend to relevant parts of the image and filter out irrelevant background clutter ensures that the extracted embeddings focus on wheat heads and their distinctive features[1]. Wheat heads in agricultural fields may be located far from each other, and their detection might require understanding long-range dependencies between them. ViT's self-attention mechanism can effectively capture such dependencies, enabling the model to establish connections between widely distributed wheat heads.

## 3. Proposed CETR deep model

A novel deep learning model CETR is proposed in this research paper for wheat head detection from farm images. CETR deep learning model performs object detection using a two-stage detection system with CenterNet and vision transformer. The major goal of CETR is to successfully evade the necessity for several manually built components, such as a non-maximum suppression approach or anchor generation, which clearly embed prior information about the task, make the process complicated, and are costly in terms of computing.

CETR architecture is a combination of the CenterNet and vision transformer deep learning models. CETR is customized to efficiently detect wheat heads from farm images. In object detection, understanding the context is crucial for accurate localization and recognition of objects, especially in cluttered or complex scenes. Its attention mechanism adaptively assigns different levels of importance to different regions of the image.

Objects of interest may be surrounded by cluttered backgrounds, such as leaves, stems, or other debris. Attention mechanism allows the model to focus on the relevant object while filtering out irrelevant background information. It allows to effectively handle occlusion scenarios by dynamically attending to visible parts of objects. ViTs are capable of processing images with variable numbers of objects effectively. This is crucial in object detection scenarios where the number of objects varies from image to image. ViTs can generalize to novel objects that might not be present in the training data, making them more versatile for real-world applications.

The aim of CETR is to leverage vision transformer's ability to capture global contextual information. CenterNet offers several advantages for extracting feature maps from farm images for wheat head detection. **Figure 1** illustrates the methodology for proposed two-stage CETR model. CenterNet deep learning architecture is used in stage-1 for extracting relevant features (feature maps) from wheat farm images. CenterNet serves as the backbone network. The feature maps generated by CenterNet are flattened and passed through the encoder-decoder mechanism of vision transformer with positional encoding. The output embeddings are generated by the decoder. The one-dimensional representation of the data is then processed by the detection heads to generate the final predictions. They predict object classes and bounding box coordinates using output from vision transformer.

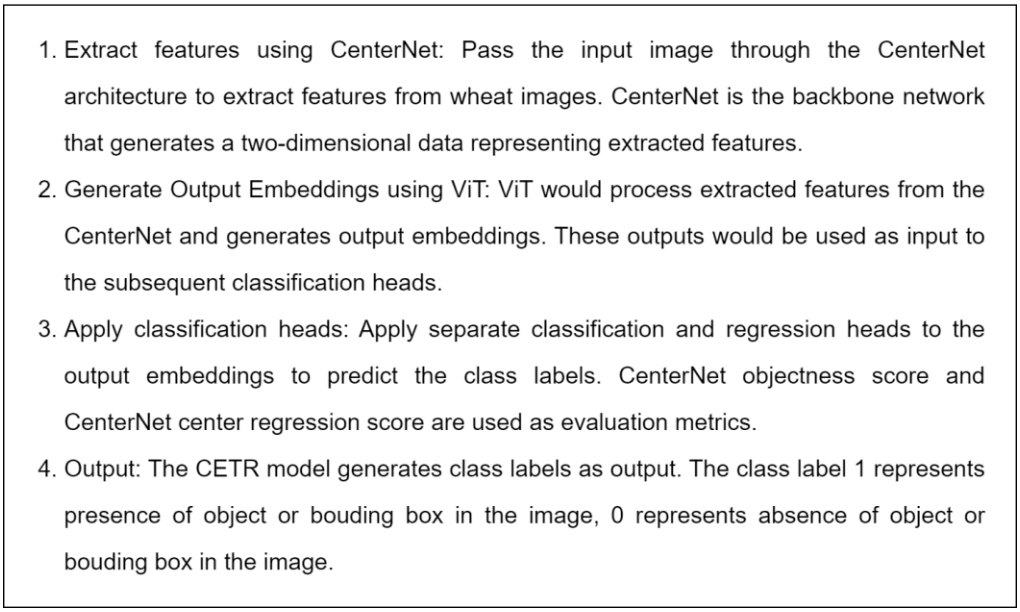


**Figure 1.** The methodology of proposed CETR for wheat seed detection.

The CETR architecture is made up consisting of three tiers (**Figure 2**).

1) CenterNet as backbone network to generate features maps from wheat farm images
2) Encoder-decoder structure of vision transformer output embeddings
3) Layer of detection heads to predict class labels and evaluate the predictions using the composite CenterNet loss function.

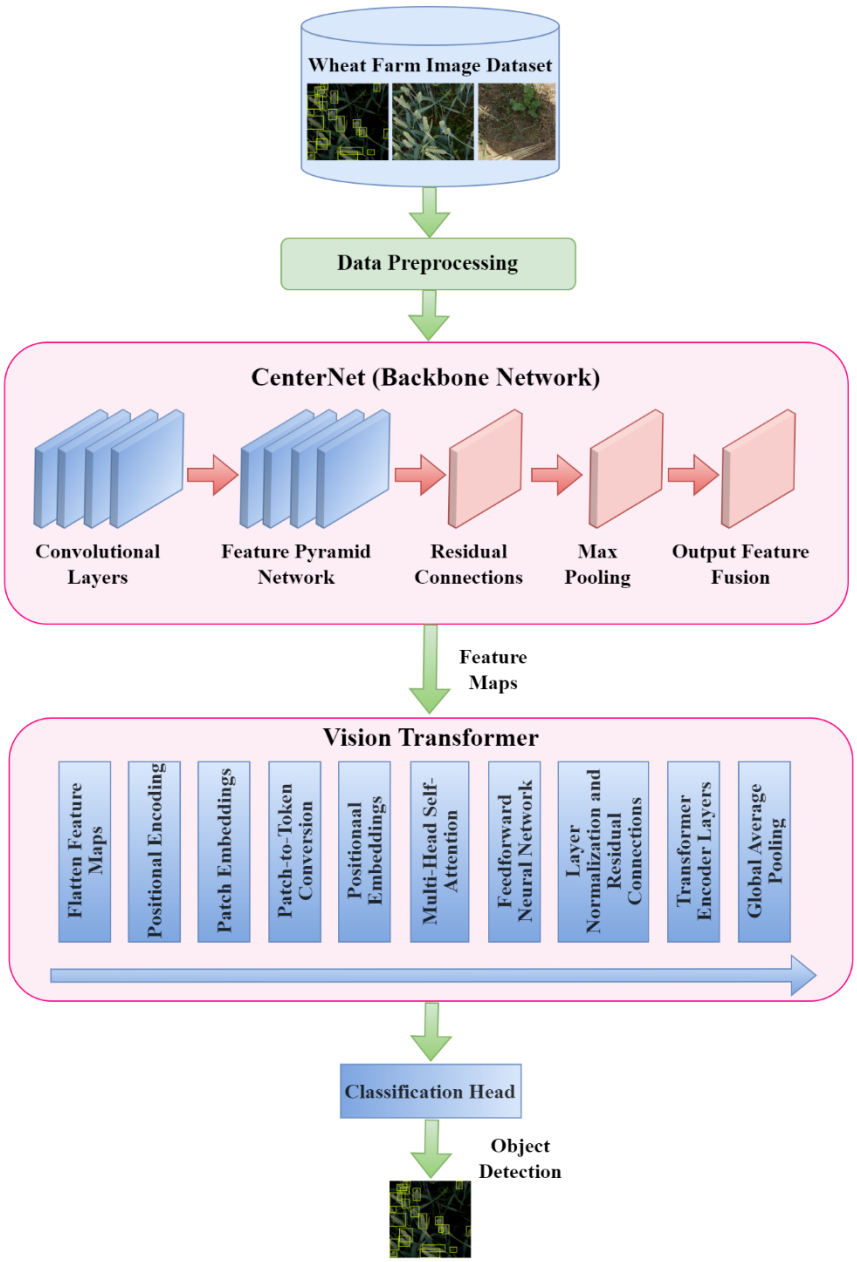


**Figure 2.** The proposed CETR architecture for wheat seed detection.

### 3.1. Stage-1: CenterNet as backbone

The CenterNet architecture is responsible for extracting feature maps from the input wheat farm images. The CenterNet architecture modules and functionalities along with its specifications are as follows.

**Input Layer:** Receive the input wheat farm images of fixed size.

**Convolutional Layers:** Four Convolutional layers are designed with filters of size 3×3 to capture different scales of features. ReLU is used to learn complex relationships between objects and backgrounds. Batch normalization improves training stability and accelerate convergence.

**Feature Pyramid Network (FPN):** Feature maps from multiple layers with different resolutions are integrated to capture both fine and coarse details. Skip connections are utilized to combine high-level and low-level features.

**Residual Connections:** Residual connections are incorporated to prevent vanishing gradients and improve gradient flow during training.

**Max Pooling:** Max pooling is applied to downsample the feature maps, reducing spatial dimensions while increasing receptive fields.

**Feature Fusion:** Feature maps are fused from different levels of the FPN to create a final feature map with enhanced representation.

**Output Feature Maps:** CenterNet architecture generates a set of feature maps that encodes rich information about the input images.

### 3.2. Stage 2: Vision transformer

Vision transformer is made up of the encoding and decoding units, as well as many transformer segments having an identical structure. A detailed description of Stage-2 of the proposed deep model, which involves using a Vision Transformer is as follows.

**Input:** Feature maps extracted by the CenterNet backbone from Stage-1. These feature maps contain encoded information about the wheat farm images.

**Flatten Feature Maps:** Feature maps in the form of 2D spatial maps are flattened to 1D vectors. Each vector represents a learned feature from a specific region of the image.

**Positional Encoding:** Positional encoding is applied to the flattened feature vectors. Positional encoding is necessary for the vision transformer to capture the spatial relationships between different features.

**Patch Embeddings:** Encoded feature vectors are divided into fixed-size patches. Each patch corresponds to a specific region of the image.

**Patch-to-Token Conversion:** Each patch is transformed into a token using a linear projection. These tokens are the input elements that the vision transformer processes.

**Positional Embeddings:** Positional embeddings are added to the tokens. to provide information about their spatial positions. This information includes both the x and y coordinates of each token within the image grid.

**Multi-Head Self-Attention:** It allows to capture global contextual information and learn relationships between different tokens.

**Feedforward Neural Network (FFN):** After self-attention, each token's representation is refined using FFN. The FFN consists of multiple fully connected layers.

**Layer Normalization and Residual Connections:** These techniques are applied to both the self-attention and FFN layers. These mechanisms stabilize training and improve gradient flow.

**Transformer Encoder Layers:** Transformer encoder layer (with self-attention and FFN components) refines the token representations and capture higher-level features.

**Global Average Pooling:** Global average pooling is performed on the token representations to aggregate information from all tokens into a fixed-length vector.

**Classification Head:** A classification head is connected to the global average-pooled vector. It uses softmax activation function. It generates class probabilities for each token, indicating the presence of a wheat head.

**Output:** The output of Stage-2 is a set of class probabilities and bounding box predictions for each token. These predictions correspond to potential wheat head locations in the input image.

The proposed CETR deep learning model exhibits the pros of CenterNet and ViT, providing an efficient and effective solution for wheat head detection. CETR has ability to effectively handle occlusion scenarios by dynamically attending to visible parts of objects. CETR can adaptively attend to visible parts of wheat heads,

even if they are partially obstructed. These capabilities makes CETR a promising approach to accurately detect wheat heads. **Figure 2** presents the proposed CETR architecture. **Figure 3** presents the proposed CenterNet architecture of stage-1.

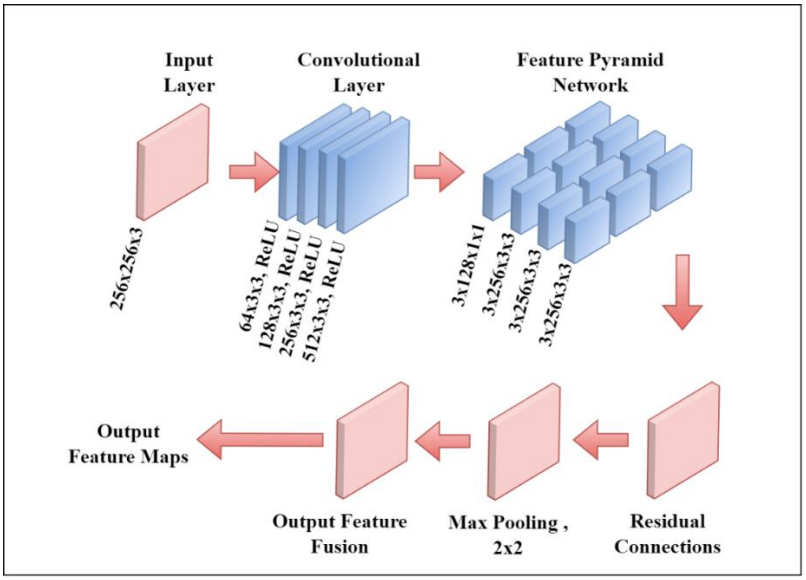


**Figure 3.** The Proposed CenterNet architecture for wheat seed detection.

### 3.3. Loss function

The CETR methodology is comprised of a loss function that enforces classification via a combination of CenterNet objectness loss and CenterNet Center Regression Loss along with a transformer-based encoder-decoder structure. CETR pursues the relationships of the objects and the larger visual background to immediately generate the final predictions, on providing a finite training object data.

#### 3.3.1. CenterNet objectness score

The CenterNet objectness loss is used to evaluate CETR. This metric measures the capability of CETR to differentiate between object centers (regions of interest) and background regions. It employs Binary Cross-Entropy (BCE) loss to penalize the misclassification of object centers. This loss function measures the difference between predicted objectness score and ground truth objectness label. This loss function helps CETR to accurately predict the presence or absence of objects at each spatial location in the output heatmap.

Let the output of the proposed model be represented as a heatmap. This heatmap is a set of binary numbers. In Heatmap, a binary digit 1 represents object centers whereas a 0 represents background. Let $H$ be a 2-dimensional heatmap of size ($L$, $W$). Let $G$ be the corresponding binary ground truth heatmap $G$ of the same size as $H$. The CenterNet objectness loss can be defined as follows.

Let $i, j$ represent any position in the heatmap $H$. Let $H_{ij}$ represent the predicted objectness score at position $(i, j)$. Let $G_{ij}$ represent the corresponding ground truth objectness label. The CenterNet objectness loss $L\_obj$ is given by Equation (1).

$$L\_obj = \frac{-1}{(L \times W)} \sum \left( G_{ij} \times \log\left(H_{ij}\right) + \left(1 - G_{ij}\right) \times \log\left(1 - H_{ij}\right) \right) \quad (1)$$

The BCE loss is computed element-wise for each position in the heatmap and then averaged across all positions. BCE loss penalizes the model for incorrect predictions of object centers (when $G_{ij} = 1$ and $H_{ij}$ is close to 0) and for incorrect predictions of background regions (when $G_{ij} = 0$ and $H_{ij}$ is close to 1). The objective of this loss function is to accurately predict the presence or absence of objects at each spatial location in the output heatmap. This helps in identifying object centers and generating bounding boxes during the object detection process in the CenterNet.

### 3.3.2. CenterNet center regression score

The CenterNet center regression loss is used to evaluate CETR. It measures the deviation of the detected object centers compared to the corresponding ground truth object centers. This loss function penalizes discrepancy between predicted and ground truth center coordinates.

Let $H$ be a 2-dimensional heatmap of size ($L$, $W$). Let $G$ be the corresponding ground truth heatmap of the same size as $H$. Each heatmap cell contains the coordinates $(x, y)$ of the ground truth center. The CenterNet center regression loss can be computed as follows:

Let $H_{ij} = (H_{x\_ij}, H_{y\_ij})$ represent the predicted center coordinates at position $(i, j)$ in the heatmap $H$. Let $G_{ij} = (G_{x\_ij}, G_{y\_ij})$ represent the corresponding ground truth center coordinates. The CenterNet center regression loss $L\_center$ is given by the Mean Squared Error (MSE) loss between predicted and ground truth center coordinates. The CenterNet center regression loss $L\_center$ is given by Equation (2).

$$L\_center = \frac{1}{(L \times W)} \sum \left( (H_{x\_ij} - G_{x_{ij}})^2 + (H_{x\_ij} - G_{x_{ij}})^2 \right) \quad (2)$$

MSE loss is computed element-wise for each position in the heatmap, and then the losses are averaged across all positions. The objective of the CenterNet center regression loss is to encourage the model to accurately predict the deviation between detected object centers and corresponding ground truth object centers. By minimizing this loss during training, the model learns to regress the centers of objects accurately. This is crucial for generating precise bounding boxes. Also, this process will efficiently localize objects during the object detection process in CenterNet.

## 4. Experimentation results and discussion

By employing CETR, researchers and agricultural practitioners can enhance the precision and efficiency of wheat head detection, resulting in improved crop management practices and heightened agricultural productivity. The strengths of CenterNet deep learning architecture in terms of accurate center point localization, robustness to occlusion makes it a promising choice for feature extraction in farm images for wheat head detection. Vision transformer plays a crucial role in generating output embeddings from feature maps extracted by CenterNet in farm images for wheat head detection. ViT's importance lies in its ability to capture global contextual information and effectively process long-range dependencies in the image, which complements the strengths of CenterNet and enhances the overall performance of the wheat head detection system.

**Figure 4** shows sample training wheat farm images. The dataset comprises of images collected with and without wheat heads. The dataset comprises of images with and without annotations of bounding boxes. Wheat image dataset is used for CETR model training[19]. CETR is trained with varying dataset sizes 1K, 2K and 3K. Batch size, number of epochs for training CETR is varied. The details are tabulated in **Table 1**. A variant of vision transformer ViT-L/16 is used as stage-2.

(a) Wheat Farm Image with Wheat Heads and with Bounding Boxes

(b) Wheat Farm Image with Wheat Heads and without Bounding Boxes

(c) Wheat Farm Image without Wheat Heads

**Figure 4.** Sample training images.

**Table 1.** Performance of CETR model.

| Dataset Size | Batch Size | Epochs | mAP | mAP50 |
|---|---|---|---|---|
| 1K | 32 | 20 | 0.8035 | 0.8155 |
| | | 50 | 0.8123 | 0.8318 |
| | 68 | 20 | 0.815 | 0.8087 |
| | | 50 | 0.8186 | 0.8046 |
| 2K | 32 | 20 | 0.8233 | 0.8164 |
| | | 50 | 0.8166 | 0.8156 |
| | 68 | 20 | 0.8123 | 0.8059 |
| | | 50 | 0.7964 | 0.8182 |
| 3432 | 32 | 20 | 0.8144 | 0.8163 |
| | | 50 | 0.8087 | 0.807 |
| | 68 | 20 | 0.7985 | 0.8006 |
| | | 50 | 0.8002 | 0.7934 |

The map, mAP50 (mean average precision at Intersection over Union threshold of 50) scores for each model are tabulated, analyzed, and conclusions are drawn based on the results. Remarkably, CETR model, achieves the highest mAP of 0.8318, surpassing the performance of CNN. CenterNet's keypoint-based approach is well-suited for detecting objects like wheat heads, which have distinctive keypoints. The two-stage detector approach using CenterNet and ViT proves more effective than employing CNN alone. Wheat heads often exhibit complex spatial arrangements in agricultural images. CenterNet, when combined with ViT, can effectively handle these complexities by detecting keypoints and considering the global context. Overall, the CETR model stands out as an efficient and compelling approach for wheat head detection.

The experimental evaluation of CETR is detailed in **Table 1**. Average precision and average recall performance of CETR is presented in **Table 2**. The COCO evaluation format provides evaluation results in terms of precision and recall, which are computed across different IoUs, areas, and maximum number of detections (maxDets). For instance, the notation [IoU = 0.50:0.95 | area = all | maxDets = 100] signifies that precision is computed over the range of IoUs from 0.5 to 0.95, with 0.05 as the step size. All detections falling within this IoU range are considered positive detections. In addition, the evaluation is performed for small, medium, and large areas, with a maximum number of detections of 100. It is noteworthy that lower IoU thresholds result in more detections being considered as true positives, thereby leading to higher precision scores.

**Table 2.** Quality metrics of CETR.

| Model | Panoptic Quality | Segmentation Quality | Recognition Quality |
|---|---|---|---|
| CETR | 44.3 | 79.3 | 53.4 |

Therefore, the highest precision score is obtained for IoU = 0.5, which has the largest number of positive detections. Conversely, when the IoU threshold is set to 0.95, fewer detections are considered as true positives, leading to lower precision scores. The IoU = 0.50:0.95 represents the average of all precisions computed across different IoUs, and hence, the precision score for this category is lower than that obtained for IoU = 0.5. The CETR model has undergone multiple experiments, and its performance has been evaluated based on accuracy scores. CETR model consistently achieved accuracy scores above 80% in most experiments. The accuracy range provided, from 79.34% to 83.18%, indicates the CETR's performance variation across different experiment setups.

**Table 3** reports the panoptic quality, segmentation quality and the recognition quality of CETR. A segmentation quality score of 79.3 indicates that the model performs relatively well in segmenting different semantic classes in the images, with a relatively high accuracy in assigning class labels to pixels. A recognition quality score of 53.4 indicates that the model's performance in instance segmentation is moderate, with scope for improvement in accurately recognizing and distinguishing individual instances of objects. The panoptic quality score is 44.3, representing the overall quality of the panoptic segmentation model on the dataset. The panoptic quality score indicates a moderate level of performance by the model. This suggests that there's potential for improvement in both the tasks of semantic segmentation and instance segmentation.

**Table 3.** Average precision and average recall performance of CETR.

| Metric | Area | IoU | Maximum Detections | Value |
|---|---|---|---|---|
| Average Precision | All | 0.50:0.95 | 100 | 0.314 |
| | All | 0.50 | 100 | 0.745 |
| | All | 0.95 | 100 | 0.259 |
| | Small | 0.50:0.95 | 100 | 0.012 |
| | Medium | 0.50:0.95 | 100 | 0.279 |
| | Large | 0.50:0.95 | 100 | 0.467 |
| Average Recall | All | 0.50:0.95 | 1 | 0.024 |
| | All | 0.50:0.95 | 10 | 0.175 |
| | All | 0.50:0.95 | 100 | 0.489 |
| | Small | 0.50:0.95 | 100 | 0.008 |
| | Medium | 0.50:0.95 | 100 | 0.438 |
| | Large | 0.50:0.95 | 100 | 0.587 |

**Table 4** and **Figure 5** show the comparative evaluation of CETR against deep models—AlexNet, VGG19, ResNet152 and MobileNet. The AlexNet model achieved an mAP50 score of 0.7160 and was the weakest model. This means that when the model predicts bounding boxes for objects (e.g., wheat heads) in the images, it has an average precision of 71.60% when considering a moderate overlap (50%) between predicted bounding boxes and ground truth. In other words, about 71.60% of the predicted bounding boxes have a good spatial agreement (intersection over union $\geq 0.50$) with the ground truth. Proposed CETR model achieved a higher mAP50 score of 0.8318. This indicates that CETR model performs better than CNN model for the undertaken task. The CETR model achieves an average precision of 83.18% when considering a 50% intersection over union threshold, implying that it has a higher accuracy in localizing wheat heads and predicting bounding boxes that closely match the ground truth.

Experimental results demonstrated that ViT effectively encoded high-dimensional feature maps extracted by CenterNet into compact and meaningful embeddings. These embeddings preserved critical information about the detected objects, allowing for accurate classification and precise bounding box regression. By using ViT to generate output embeddings, the proposed CETR model was able to refine the localization of detected object centers, leading to improved accuracy in bounding box predictions. ViT is capable of learning global patterns and representations from large datasets. CETR leveraged this capability of ViT to generalize the knowledge gained from diverse agricultural farm images. This led to improved performance of CETR on new, unseen data.

**Table 4.** Comparative evaluation of the proposed model.

| Model | mAP50 | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|---|
| Alexnet | 0.7160 | 78.0 | 71.0 | 79.6 | 89.5 |
| VGG19 | 0.7638 | 80.7 | 75.4 | 78.7 | 81.7 |
| ResNet152 | 0.7702 | 86.1 | 85.8 | 87.4 | 86.3 |
| MobileNet | 0.7854 | 81.2 | 77.9 | 84.2 | 89.0 |
| CETR | 0.8318 | 93.1 | 83.9 | 87.2 | 88.6 |

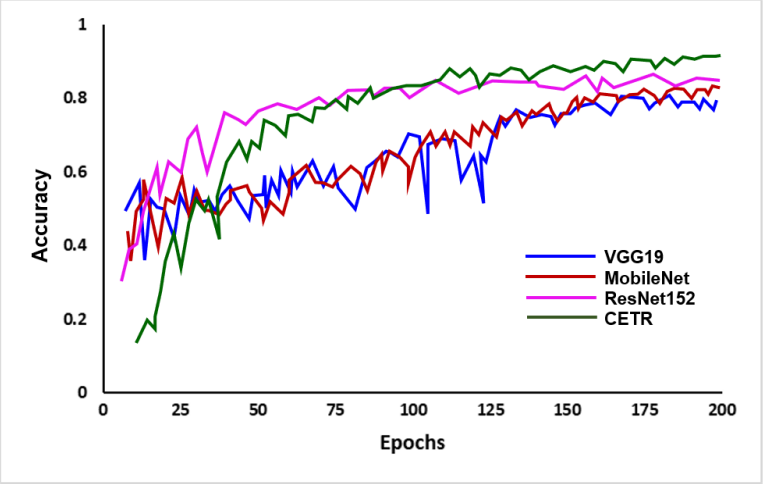


**Figure 5.** Comparative evaluation of CETR model.

## 5. Conclusion

The proposed CETR deep learning model proves to be a highly effective and promising approach for wheat head detection from agricultural farm images. By adopting a transformer-based architecture, our model can efficiently process agricultural field images and capture spatial relationships between wheat heads, enabling effective detection in challenging conditions. Through its innovative combination of CenterNet and vision transformer, the CETR model leverages the strengths of both architectures, achieving outstanding results. Usage of vision transformer in the wheat head detection pipeline enhanced the understanding of global context by capturing long-range dependencies, and effectively encoded features from CenterNet's extracted feature maps. This fusion of CenterNet and ViT leveraged the strengths of both approaches, leading to more accurate and robust wheat head detection in agricultural farm images. The model performed relatively well in semantic segmentation with a score of 79.3, but its performance in instance segmentation is relatively moderate, obtaining a score of 53.4. The PQ score is 44.3, representing the overall quality of the panoptic segmentation model on the dataset. These scores provide insights into the strengths and limitations of CETR and thus can suggest areas for potential improvement. The higher mAP value 0.8318 for CETR compared to 0.7160 of CNN indicates that the CETR provides more effective wheat head detection in agricultural images. It achieves a higher precision in predicting bounding boxes that align well with the ground truth, resulting in more accurate and reliable wheat head detection. In conclusion, this research introduces CETR, a CenterNet-Vision Transformer model, which addresses the challenges of wheat head detection in agricultural farm images. The model's high accuracy, robustness, and efficiency make it a valuable tool for advancing crop yield estimation and agricultural decision-making, thus contributing to sustainable and efficient agricultural practices.

## Author contributions

Conceptualization, KGS, GS; methodology, KGS, GS, RK; software, PKK, NK; validation, RK, ERA; formal analysis, KGS, ERA; investigation, KA, ERA; resources, KA; data curation, PKK; writing-original

draft preparation, GS; writing—review and editing, NK; visualization, GS; supervision, RK. All authors have read and agreed to the published version of the manuscript.

## Conflict of interest

The authors declare no conflict of interest

## References


1. Khaki S, Safaei N, Pham H, et al. WheatNet: A lightweight convolutional neural network for high-throughput image-based wheat head detection and counting. Neurocomputing. 2022, 489: 78-89. doi: 10.1016/j.neucom.2022.03.017
2. Khan S, Naseer M, Hayat M, et al. Transformers in Vision: A Survey. ACM Computing Surveys. 2022, 54(10s): 1-41. doi: 10.1145/3505244
3. Duan K, Bai S, Xie L, et al. Centernet: Keypoint triplets for object detection. Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019; IEEE, pp. 6569-6578.
4. Han K, Wang Y, Chen H, et al. A Survey on Vision Transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023, 45(1): 87-110. doi: 10.1109/tpami.2022.3152247
5. Dosovitskiy A, Beyer L, Kolesnikov A et al. An image is worth 16×16 words: Transformers for image recognition at scale. arXiv 2020, arXiv:2010.11929.
6. Nennuri R, Kumar RH, Prathyusha G, et al. A Multi-Stage Deep Model for Crop Variety and Disease Prediction. 14th International Conference on Soft Computing and Pattern Recognition 2023, 48: 52-59. doi: 10.1007/978-3-031-27524-1_6
7. Charan NS, Narasimhulu T, Bhanu Kiran G, et al. Solid Waste Management using Deep Learning 14th International Conference on Soft Computing and Pattern 2023, 648: 44-51. doi: 10.1007/978-3-031-27524-1_5
8. Shereesha M, Hemavathy C, Teja H, et al. Precision Mango Farming: Using Compact Convolutional Transformer for Disease Detection. 13th International Conference on Innovations in Bio-Inspired Computing and Applications 2023, 649: 458-465. doi: 10.1007/978-3-031-27499-2_43
9. Balakrishna N, Sunitha G, Karthik A, et al. Tomato Leaf Disease Detection Using Deep Learning: A CNN Approach. International Conference on Data Science, Agents & Artificial Intelligence 2022, IEEE.
10. Sudarsana Murthy D. An Investigative Study of Shallow, Deep and Dense Learning Models for Breast Cancer Detection based on Microcalcifications. 2022 International Conference on Data Science, Agents & Artificial Intelligence. pp. 1-6.
11. Thatikonda SS. Vision Transformer based ResNet Model for Pneumonia Prediction. 4th International Conference on Electronics and Sustainable Communication Systems 2023, IEEE.
12. Kumar LA, Renuka DK, Rose SL, et al. Deep learning based assistive technology on audio visual speech recognition for hearing impaired. International Journal of Cognitive Computing in Engineering. 2022, 3: 24-30. doi: 10.1016/j.ijcce.2022.01.003
13. Carion N, Massa F, Synnaeve G, et al. End-to-End Object Detection with Transformers. Lecture Notes in Computer Science. 2020, 213-229. doi: 10.1007/978-3-030-58452-8_13
14. Girshick R, Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision 2015, pp. 1440-1448. doi: 10.1109/iccv.2015.169
15. Henry EU, Emebob O, Omonhinmin CA. Vision transformers in medical imaging: A review. arXiv 2022, arXiv:2211.10043.
16. Luo J, Li B, Leung C, A Survey of Computer Vision Technologies In Urban and Controlled-environment Agriculture. arXiv 2022, arXiv:2210.11318.
17. Wu S, Sun Y, Huang H. Multi-granularity Feature Extraction Based on Vision Transformer for Tomato Leaf Disease Recognition. 3rd International Academic Exchange Conference on Science and Technology Innovation 2021. IEEE, pp. 387-390.
18. Li X, Fan W, Wang Y et al. Detecting Plant Leaves Based on Vision Transformer Enhanced YOLOv5 3rd International Conference on Pattern Recognition and Machine Learning 2022, Springer, pp. 32-37.
19. David E, Madec S, Sadeghi-Tehran PH 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. 2020. doi: 10.34133/2020/3521852