Int. J. Adv. Eng. Pure Sci. 2025, 37(2): <133-143>
DOI: 10.7240/jeps.1623086

RESEARCH ARTICLE / ARAŞTIRMA MAKALESİ

A CBAM-Enhanced UNetFormer for Semantic Segmentation of Wheat Yellow-Rust Disease Using Multispectral Remote Sensing Images

İrem ÜLKÜ [1]

[1]*Ankara University, Department of Computer Engineering, 06830, Ankara, Türkiye*

Abstract

This study focuses on the problem of wheat yellow-rust disease caused by climate change and incorrect farming methods. Early detection of the disease, which manifests as yellow-orange spores on wheat leaves, is crucial for mitigating issues such as reduced crop yield, increased pesticide use, and environmental harm. Current CNN-based semantic segmentation models focus mainly on processing local pixels, which can be insufficient for large areas. This study proposes a novel version of the UNetFormer architecture, enhancing the CNN-based encoder with CBAM modules while utilizing a Transformer-based decoder to address the limitations of current approaches. Specifically, the model incorporates a Convolutional Block Attention Module (CBAM) to refine feature extraction along spatial and channel axes. CBAM modules allow the network to prioritize meaningful features, particularly near-infrared (NIR) wavelength reflections critical for detecting wheat yellow-rust. The proposed UNetFormer2 model effectively captures long-range dependencies in multispectral remote sensing images to improve disease detection across large agricultural areas. Specifically, the model achieves an IoU improvement of 2.1% for RGB, 4.6% for NDVI, and 3% for NIR compared to the baseline UNetFormer model. This work aims to improve wheat yellow-rust disease monitoring efficiency and contribute to more sustainable agricultural practices by reducing unnecessary pesticide application.

Keywords: Semantic segmentation, Remote sensing, Multispectral images, Wheat yellow-rust disease

I. INTRODUCTION

Wheat yellow rust is a common plant disease caused by climate changes and inappropriate agricultural management strategies [1]. Generally, the traditional management tool for wheat yellow rust disease is to implement chemical methods [2]. The current approach of chemical control, applied regardless of the disease's current status, leads to excessive pesticide use, causing environmental impacts such as groundwater pollution and the risk of pesticide residues in agricultural products. As the disease progresses, it manifests as yellow-orange spores on wheat leaves due to physical and chemical changes, such as decreased chlorophyll content and water levels in the leaves [3]. Multispectral bands are crucial in detecting the yellow-rust spores by providing discriminative information [4]. Remote sensing systems via satellite or aerial vehicles can dynamically monitor plant stress over large areas [5]. Convolutional neural networks (CNNs) applied to multispectral remote sensing images are attracting attention in the literature to detect wheat yellow rust disease [6-8]. However, the processing of local pixels or small regional information in these studies is often insufficient, especially for large agricultural areas, making it crucial to capture long-range dependencies [9].

Based on these discussions, this study proposes a novel version of UNetFormer that includes a CNN-based encoder and a Transformer-based decoder to capture long-range dependencies [10]. In the UNetFormer architecture, the ResNet18-based CNN encoder integrates channel and spatial information to extract meaningful features. This study proposes a Convolutional Block Attention Module (CBAM) [11] used in UNetFormer design to emphasize meaningful features along the channel and spatial axes. Indicators of wheat yellow rust, such as leaf color changes and reduced chlorophyll content, are detected using near-infrared (NIR) wavelength reflection [12, 13]. The proposed CBAM module enables UNetFormer to automatically learn the importance of the NIR wavelength for disease detection and suppress irrelevant wavelengths. Additionally, using the CBAM module allows the detection of even small lesions appearing in the early stages of the disease, thus enhancing the information flow efficiency within the UNetFormer architecture.

Corresponding Author: IREM ÜLKÜ, **Tel:** 0312 203 33 00-1756, **E-mail:** irem.ulku@ankara.edu.tr

Submitted: 19.01.2025, **Revised:** 28.03.2025, **Accepted:** 17.04.2025

The structure of this paper is in the following manner: Section II presents a comprehensive review of existing semantic segmentation studies using remote sensing images. Section III introduces the core components of the proposed architecture, while Section IV provides information about the WYR image set used in this study. Section V covers implementation details and evaluation metrics, followed by visualization and analysis of the test results in Section VI. Finally, Section VII discusses the study's conclusions.

II. RELATED WORK

This section presents the most common semantic segmentation architectures. Early approaches use CNN models, but recent efforts also utilize transformers to yield better performance.

2.1. Convolutional Neural Networks (CNNs)

U-Net, with its symmetric expansion-contraction paths architecture, offers a deep network capable of capturing context while providing precise localization [14]. SegNet, on the other hand, performs pixel-wise semantic segmentation with an encoder-decoder architecture, using pooling indices obtained from the encoder to enable nonlinear upsampling in the decoder, thus achieving memory- and computation-efficient performance [15]. DeepLabv3+ combines atrous spatial pyramid pooling, which models contextual information at various scales, with a decoder module that enhances object boundaries, forming an efficient architecture [16]. BiSeNet is a real-time model balancing speed and segmentation performance by combining a spatial path that provides high-resolution location information with a context path that captures a large receptive field [17]. DFANet, an efficient CNN architecture, starts with a lightweight backbone and employs cascades of sub-networks and sub-stages to aggregate discriminative features, aiming to reduce parameter count while balancing speed and segmentation performance through multi-scale feature propagation [18]. D-LinkNet, designed for road extraction, combines dilated convolution and a pre-trained encoder while retaining the computational efficiency of the LinkNet architecture [19].

There are also many CNN-based studies on the pixel-wise classification of aerial imagery. For instance, LANet, a CNN-based architecture, improves performance in the semantic segmentation of remote sensing images by combining high-level semantic understanding with low-level location details [20]. Another CNN-based architecture, AFNet, performs effectively in complex urban landscapes with a Scale Feature Attention Module (SFAM) to detect objects of different sizes [21].

U-Net, with its symmetric expansion-contraction paths architecture, offers a deep network capable of capturing context while providing precise localization [14]. SegNet, on the other hand, performs pixel-wise semantic segmentation with an encoder-decoder architecture, using pooling indices obtained from the encoder to enable nonlinear upsampling in the decoder, thus achieving memory- and computation-efficient performance [15]. DeepLabv3+ combines atrous spatial pyramid pooling, which models contextual information at various scales, with a decoder module that enhances object boundaries, forming an efficient architecture [16]. BiSeNet is a real-time model balancing speed and segmentation performance by combining a spatial path that provides high-resolution location information with a context path that captures a large receptive field [17]. DFANet, an efficient CNN architecture, starts with a lightweight backbone and employs cascades of sub-networks and sub-stages to aggregate discriminative features, aiming to reduce parameter count while balancing speed and segmentation performance through multi-scale feature propagation [18]. D-LinkNet, designed for road extraction, combines dilated convolution and a pre-trained encoder while retaining the computational efficiency of the LinkNet architecture [19].

2.2. Transformers

The Vision Transformer (ViT), solely based on transformers, applies the transformer architecture to sequences of image patches, achieving excellent results on various image recognition benchmarks [22]. A high-performing vision transformer, DeiT [23], suggests a distillation strategy that leverages tokens, using a CNN model as the teacher on the ImageNet-1k dataset [24]. Swin Transformer is a hierarchically organized transformer architecture that extracts representations through shifted windows [25]. SegFormer combines Transformers with efficient multilayer perceptron (MLP) decoders, offering robust semantic segmentation that operates without positional encoding [26]. Segmenter, a transformer model, uses contextual information at the image patch level to achieve label consensus and offers high performance by modeling the global context from the initial layer to produce class labels from the corresponding embedded outputs of image patches [27]. The SEgmentation TRansformer (SETR) architecture represents the image as a sequence of patches and modeling the global context at each layer [28]. TransUNet uses transformers to model long-range dependencies while capturing high-resolution local details with the U-Net architecture [29]. CCTUnet combines CNN and transformer to enhance the preservation of edge details in high-resolution input images and capture long-range dependencies [30].

Transformer-based architectures are also increasingly used with CNN-driven methods for the semantic understanding of remote sensing images, demonstrating superior performance. For instance,

EMRT, a hybrid architecture designed for high-resolution remote sensing images, aims to enhance multi-scale representation learning by combining the strength of CNNs in deriving local features with the capability of transformers in learning global patterns [31]. CTCFNet, a fused CNN-transformer structure for remote sensing images, improves segmentation accuracy by integrating local and global information to overcome occlusions of objects and surface features at different scales [32]. One study [33] employs a Swin Transformer-based encoder and CNN structures in the decoder to better model long-range spatial dependencies. The CMTFNet model extracts and combines local and global contextual information at multiple scales for remote sensing images [34]. ST-U-Net model strengthens the feature representation ability by embedding the Swin transformer within the U-Net structure for the pixel-wise classification of remote sensing imagery [35]. CSTUNet utilizes a two-stream encoder comprising a CNN-based primary encoder and a Swin transformer-assisted secondary encoder [36]. UNetFormer [10] is designed for real-time segmentation of urban scene images in remote sensing and uses a lightweight ResNet18 encoder, offering an effective attention mechanism that jointly processes global and local information. This study proposes to adaptively focus on essential features in both the channel and spatial dimensions by utilizing the CBAM module in the UNetFormer design for the semantic segmentation of wheat yellow rust disease.

III. MATERIALS AND METHODS

Figure 1 illustrates the proposed architecture, where each feature map obtained from the CNN-based encoder is passed through a separate CBAM module to enhance adaptive feature improvement. The CBAM modules generate attention maps along the channel and spatial dimensions. Then, these complementary attentions are multiplied with the input feature maps to serve as the refined output. This process allows the model to learn to emphasize features that are supposed to ease the detection of wheat yellow-rust objects while suppressing irrelevant ones. Finally, the transformer-based decoder generates the segmentation output. The following sections explain the details of the architecture.

3.1. Encoder

The encoder structure, consisting of four blocks and shown in Figure 1, is designed to extract multi-scale semantic features using the pre-trained ResNet18 model [37]. ResNet18, chosen for its relatively lightweight architecture with four blocks, performs downsampling by a factor of 2 at each block. Assuming that R_k represents the output of each ResNet block, where k represents the block layer level, there are four levels in total. The resulting output is A_k when using a CBAM module for each block k. Therefore, each ResNet block output processed through the CBAM module is as follows:

$$A_k = CBAM(R_k) \quad (1)$$

Each feature map A_k produced by the CBAM is combined with the corresponding feature maps in the decoder using a weighted sum with a 1 × 1 convolution. The final fused feature map, denoted as FF_k, is obtained as follows:

$$FF_k = \alpha . A_k + (1 - \alpha) . GLF_k \quad (2)$$

where α is the weighting coefficient, and GLF_k represents the decoder feature map at the corresponding level.

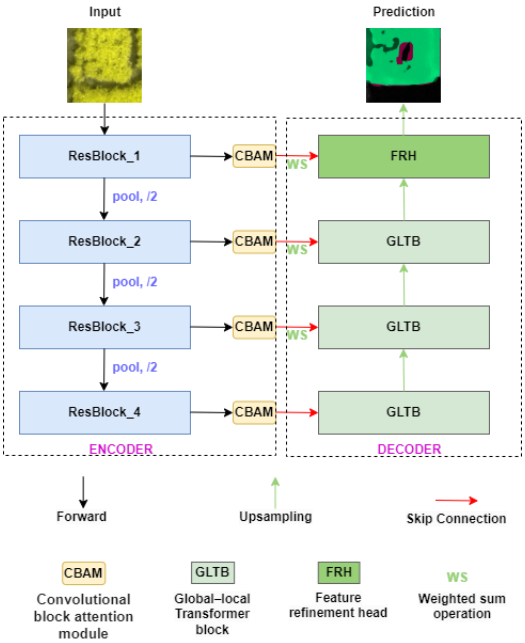


Figure 1. Overview of the proposed UNetFormer2 architecture

3.1.1. CBAM modules

For each feature map $R_k \in R^{C_k \times H_k \times W_k}$ obtained from the ResNet18 blocks, a separate CBAM module processes it to produce a channel attention map $M_c \in R^{C_k \times 1 \times 1}$ and a spatial attention map $M_s \in R^{1 \times H_k \times W_k}$. Figure 2 illustrates the CBAM module operations, which are detailed below:

$$A'_k = M_c(R_k) \otimes R_k \quad (3)$$

$$A_k = M_s(A'_k) \otimes A'_k \quad (4)$$

where $\otimes$ denotes element-wise multiplication, and A_k is the final output obtained for each level k.

The CBAM obtains channel attention by compressing the spatial dimension using the average-pooled features R^c_{avg} and max-pooled features R^c_{max}. Figure 2 shows how these feature maps pass through a multi-layer perceptron (MLP) network with a single hidden

layer, which then computes the channel attention map $M_c \in R^{C_k \times 1 \times 1}$ by combining the output features element-wise as follows:

$$M_c(R_k) = \sigma\left(MLP(R^c_{avg}) + MLP(R^c_{max})\right) = \sigma\left(W_1\left(W_0(R^c_{avg})\right) + W_1(W_0(R^c_{max}))\right) \quad (5)$$

where $W_0 \in R^{C_k/r \times C_k}$ and $W_1 \in R^{C_k \times C_k/r}$ are the weights of the MLP network. These weights remain shared across all inputs. The final sigmoid function appears as σ, and a ReLU activation function applies after W_0.

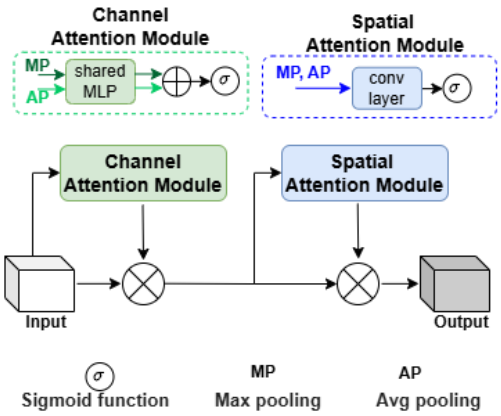


Figure 2. CBAM structure

To compute spatial attention, as shown in Figure 2, the outputs $R^s_{avg} \in R^{1 \times H_k \times W_k}$ and $R^s_{max} \in R^{1 \times H_k \times W_k}$, obtained by applying average pooling and max pooling operations along the channel axis, are combined. A convolutional layer with a filter size of 7 × 7, denoted as $f^{7\times7}$, is then applied to this combined feature map to produce the spatial attention map $M_s \in R^{H_k \times W_k}$ as follows:

$$M_s(R_k) = \sigma\left(f^{7\times7}\left([R^s_{avg}; R^s_{max}]\right)\right) \quad (6)$$

3.2. Decoder

The decoder section consists of three global-local Transformer blocks and one feature enhancement head, as shown in Figure 1. The following sections explain all these processes in detail.

3.2.1. Global-local transformer block (GLTB)

Figure 3 illustrates the global-local attention structure that the GLTB block uses. It consists of parallel local and global branches that process the input simultaneously. Each feature map $A_k \in R^{C_k \times H_k \times W_k}$ obtained from the CBAM blocks undergoes a 1 × 1 convolution operation, adjusting the channel size to a fixed value of $C = 64$. The resulting feature map $A''_k \in R^{B \times C \times H_k \times W_k}$ serves as input to both the local and global branches, with a batch size of B. The local branch extracts both small and large-scale local features from the 2D feature map A''_k by passing it through 1 × 1 and 3 × 3 convolution operations separately. After applying batch normalization (BN) to each output, the branch combines the results.

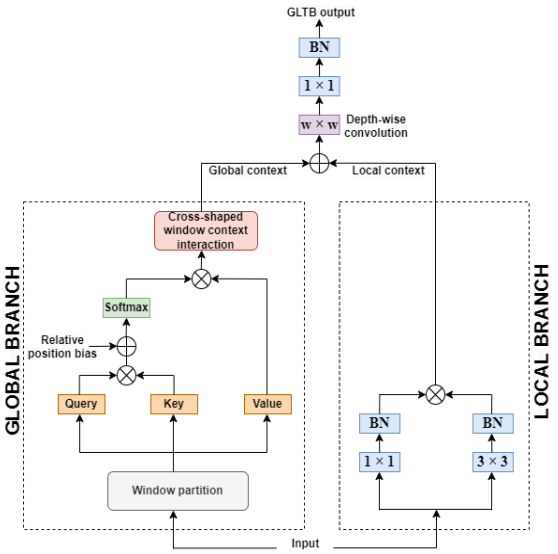


Figure 3. Illustration of GLTB blocks

The global branch aims to capture global context by applying window-based multi-head self-attention to the input $A''_k \in R^{B \times C \times H_k \times W_k}$, as shown in Figure 4. First, a 1 × 1 convolution is applied to the feature map A''_k to triple the channel size. When the dimension becomes $R^{B \times 3C \times H_k \times W_k}$, it provides more feature information related to each pixel. Next, the window partitioning operation sets the window size to $w \times w$, resulting in an output dimension of $R^{B \times 3C \times (H_k/w) \times (W_k/w) \times w \times w}$. The process resizes the information in each $w \times w$ window into a 1D array suitable for the attention mechanism.

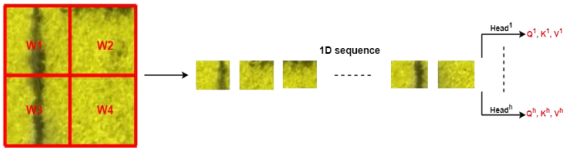
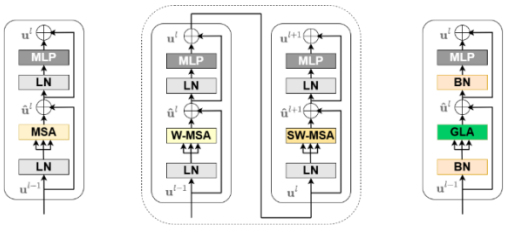


Figure 4. Illustration of window partitioning



(a) **(b)** **(c)**

Figure 5. (a) Transformer block structure. (b) Swin Transformer block structure. (c) UNetFormer Transformer block structure.

When the attention mechanism uses h multi-heads, it defines the dimension for each head as $R^{B \times 3 \times h \times (\frac{C}{h}) \times (H_k/w) \times (W_k/w) \times (w \times w)}$. The 1D arrays

obtained for each head convert into query (Q), key (K), and value (V) vectors.

The original Swin Transformer architecture uses a window shifting mechanism (Figure 5) to slightly shift the windows in consecutive Swin Transformer layers, allowing information interaction between neighboring windows; however, this process also increases the computational load. In contrast, UNetFormer employs a cross-shaped window context interaction module to capture intra-window and inter-window long-range relationships within a transformer layer at a lower computational cost. UNetFormer Transformer block is illustrated in Figure 5(c).

The cross-shaped window context interaction module, visualized in Figure 6, combines two feature maps generated by horizontal and vertical average pooling layers to obtain global context. The dependency of any point $p_1^{(m,n)}$ in Window 1 on the point $p_2^{(m+w,n)}$ in Window 2 is calculated as follows:

$$P_1^{(m,n)} = \frac{\sum_{i=0}^{w-m-1} P_1^{(m+i,n)} + \sum_{j=0}^{m} P_2^{(m+w-j,n)}}{w} = \frac{\sum_{i=0}^{w-m-1} D_i\left(P_1^{(m,n)}\right) + \sum_{j=0}^{m} D_j\left(P_2^{(m+w,n)}\right)}{w} \quad (7)$$

$$P_1^{(m+i,n)} = D_i\left(P_1^{(m,n)}\right) \quad (8)$$

$$P_2^{(m+w-j,n)} = D_j\left(P_2^{(m+w,n)}\right) \quad (9)$$

where w denotes the window size and D represents the self-attention computation used to model the interactions of pixel pairs within a window. For example, the interaction between a point $p_1^{(m,n)}$ from Window 1 and any other point $p_1^{(m+i,n)}$ within the same window is expressed by equation (8).

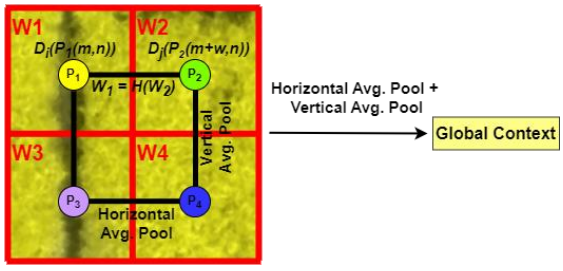


Figure 6. Illustration of cross-shaped window context interaction module operations

Equation (9) calculates the interaction between a point $p_2^{(m+w,n)}$ within Window 2 and any other point $P_2^{(m+w-j,n)}$ in this window. The horizontal relationship between the point $p_1^{(m,n)}$ in Window 1 and the point $p_2^{(m+w,n)}$ in Window 2 is established by equation (7). The same equation also enables the calculation of the vertical relationship between the point $p_1^{(m,n)}$ in Window 1 and a point $p_3^{(m,n+w)}$ in Window 3. This process is beneficial to extract long-range dependencies between any two windows and to construct the global context.

3.2.2. Feature refinement head (FRH)

The FRH structure aims to enhance segmentation accuracy by closing the information gap between shallow and deep feature maps through a refined merging process, using the combined feature map FF_k shown in equation (2) as input. As illustrated in Figure 7, a global average pooling layer creates a channel-wise attention map $C \in R^{1\times1\times c}$ with c channel dimension. The reduce & expand operation applied at this stage consists of two 1 × 1 convolution layers that reduce the channel size by a factor of 4 and then expand it back to its original size.

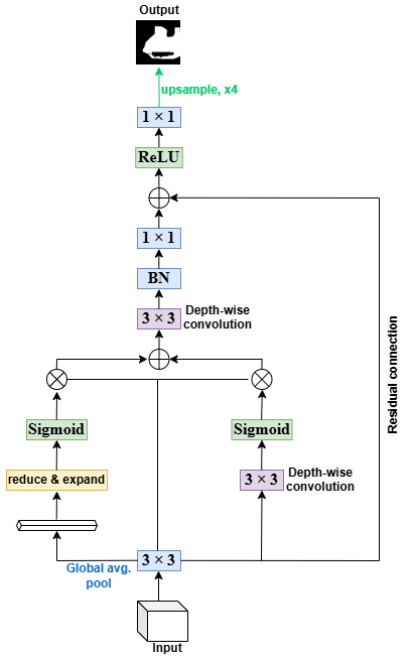


Figure 7. Illustration of FRH operations

Meanwhile, a depth-wise convolution applies to produce a spatial-wise attention map $S \in R^{h\times w\times1}$ for h and w spatial resolutions. The feature maps obtained from the two parallel paths are combined. Adding the residual connection at this part is a powerful tool to prevent potential information loss. Subsequently, a 1 × 1 convolution layer and an upsampling operation are applied. As a result, the process produces a segmentation map that encodes detailed spatial information and semantic meaning.

IV. RESULTS

The experiments evaluate the proposed UNetFormer2 model on the wheat yellow rust disease image set and compare it with state-of-the-art (SOTA) semantic segmentation models. This section describes the implementation details, image set, and experimental results.

4.1. Implementation Details

The models are trained with an initial learning rate of $5 * 10^{-5}$, using a strategy where the learning rate decreases by 9% every ten epochs on an NVIDIA Quadro RTX 5000 GPU. The selected optimizer is the Adam optimizer with a momentum value of 0.9 and a batch size of 8. The experiments run for 70 epochs, with 5-fold cross-validation employed. All images, originally sized 1336 × 2991 pixels, are divided into 224 × 224 pixel-sized patches with overlap, resulting in 1299 patches. The image set, which is not publicly available, has a sample distribution of 72% for training, 20% for testing, and 8% for validation, with the split determined based on the convention established in [8]. Hyperparameters are manually tuned based on prior experience and experimental results.

4.2. The Wheat Yellow-Rust (WYR) Image Set

The WYR image set comes from field experiments conducted in 2019 at the Caoxinzhuang Experimental Station in the Yangling region of China to monitor yellow-rust disease in wheat plants [9]. The experiments focus on Xiaoyan 22 wheat, a variety susceptible to this disease. Yellow-rust inoculates wheat plots measuring 2 m × 2 m.

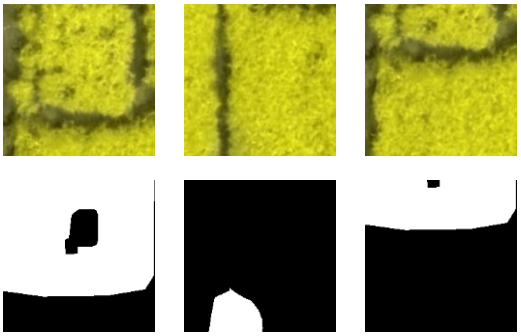


Figure 8. Examples of the WYR image set. The first row shows the input images, which are visualized using three selected bands from the multispectral dataset mapped to the RGB channels. The second row shows the ground-truth binary masks for wheat yellow rust.

A DJI Matrice 100 (M100) Quad-copter and a RedEdge multispectral camera capture images at a flight height of 20 meters, achieving a spatial resolution of 1.3 cm/pixel. The RedEdge camera captures images in the blue, green, red, red-edge, and NIR bands. Manual labeling identifies pixels representing yellow-rust disease in the images. Selected images from the WYR image set appear in Figure 8.

4.3. Performance Evaluation

IoU (Intersection over Union) and F_1 score are the semantic segmentation evaluation metrics used in the experiments. IoU shows the intersection between the ground truth and predicted pixels. The F_1 score is the harmonic mean of recall and precision calculations. Precision indicates the percentage of correctly predicted yellow-rust pixels among all pixels identified as yellow-rust. Recall measures how many of the actual yellow-rust pixels are correctly identified by the model. IoU and F_1 score calculations are as follows:

$$IoU = \frac{TP}{(TP + TN + FP)} \quad \textbf{(10)}$$

$$F_1 = \frac{(2 \times Precision \times Recall)}{(Precision + Recall)} \quad \textbf{(11)}$$

$$Precision = \frac{TP}{(TP + FP)} \quad \textbf{(12)}$$

$$Recall = \frac{TP}{(TP + FN)} \quad \textbf{(13)}$$

where TP represents the yellow-rust pixels predicted correctly, FP represents the pixels incorrectly predicted as yellow-rust, FN represents the missed yellow-rust pixels, and TN represents the pixels correctly predicted as not yellow-rust.

4.4. Experimental results

The experiments compare the proposed UNetFormer2 with various SOTA semantic segmentation models on the WYR image set. The selected SOTA models are U-Net [14], SegNet [15], DLinkNet [19], NestedU-Net [38], DeepLabV3 [16], BiSeNet [17], DFANet [18], UNetFormer [10]. The quantitative results for RGB, NDVI, and NIR images are provided in Tables 1, 2, and 3, respectively.

Table 1. Quantitative comparison of the WYR test set results against SOTA models by using RGB images. The bold values indicate the best performance.

Architectures	RGB Images	
	IoU	F_1
U-Net	0.521±0.294	0.647±0.333
SegNet	0.545±0.347	0.620±0.372
DLinkNet	0.608±0.296	0.702±0.295
NestedU-Net	0.615±0.264	0.718±0.242
DeepLabV3	0.426±0.315	0.527±0.321
BiSeNet	0.435±0.318	0.535±0.319
DFANet	0.489±0.321	0.587±0.320
UNetFormer	0.664±0.241	0.769±0.203
UNetFormer2 (Proposed)	**0.685±0.235**	**0.784±0.210**

Table 1 shows the experimental test results for semantic segmentation on the WYR image set using RGB images. UNetFormer2 surpasses the original

UNetFormer with the best performance among all the SOTA models by a 2.1% gain in IoU on the test set. UNetFormer2 consistently outperforms lightweight semantic segmentation models, i.e., BiSeNet and DFANet, by 25% and 19.6%, respectively. In addition, against the CNN-based NestedU-Net, UNetFormer2 produces a promising gain of 7%, yielding a new state-of-the-art for RGB images with 0.685 IoU.

Table 2. Quantitative comparison of the WYR test set results against SOTA models by using NDVI images. The bold values indicate the best performance.

Architectures	NDVI Images	
	IoU	F_1
U-Net	0.502±0.355	0.582±0.353
SegNet	0.469±0.366	0.545±0.378
DLinkNet	0.472±0.298	0.575±0.304
NestedU-Net	0.560±0.292	0.653±0.279
DeepLabV3	0.465±0.311	0.565±0.308
BiSeNet	0.497±0.297	0.605±0.299
DFANet	0.448±0.323	0.546±0.330
UNetFormer	0.515±0.320	0.615±0.313
UNetFormer2 (Proposed)	**0.561±0.290**	**0.666±0.282**

In Table 2, the experiments inspect the effect of vegetation indices to find out which model is more promising for the semantic segmentation of NDVI images. Compared to UNetFormer, the proposed model performs more accurately in NDVI images and improves the IoU and F_1 score metrics. It achieves 4.6% in IoU, which outperforms the original UNetFormer model. Compared to the NestedU-Net, one of the best-performing models for NDVI images, the proposed model obtains the F_1 score improvement of 1.3% on the test set. This result demonstrates that the output feature maps produced by UNetFormer2 effectively emphasize more meaningful features semantically related to the wheat yellow-rust regions. Compared to U-Net, UNetFormer2 achieves 5.9% and 8.4% gains on the IoU and F_1 performances, respectively.

The experimental test results in Table 3 analyze the effectiveness of UNetFormer2 by extending the spectral channels to NIR wavelengths. Compared to UNetFormer, the proposed model demonstrates a 3% improvement in IoU on NIR images, effectively capturing the wheat yellow-rust regions. UNetFormer2 increases IoU by 9.5% compared to DLinkNet, confirming that it can predict the wheat yellow-rust pixels more in line with ground truth. Moreover, UNetFormer2 improves IoU on NIR images by 14.4% compared to BiSeNet, the classical real-time model, while increasing the F_1 score by 12.8%. The proposed model also achieves promising performance among the classical CNN-based architectures like U-Net, SegNet, DLinkNet, NestedU-Net, and DeepLabV3. The proposed model with the ResNet18 backbone outperforms all other SOTA models and sets new state-of-the-art with an IoU score of 0.718 on the NIR test set.

Table 3. Quantitative comparison of the WYR test set results against SOTA models by using NIR images. The bold values indicate the best performance.

Architectures	NIR Images	
	IoU	F_1
U-Net	0.522±0.260	0.645±0.235
SegNet	0.466±0.312	0.562±0.322
DLinkNet	0.623±0.261	0.728±0.235
NestedU-Net	0.677±0.194	0.786±0.152
DeepLabV3	0.576±0.286	0.684±0.264
BiSeNet	0.574±0.248	0.694±0.217
DFANet	0.481±0.285	0.594±0.276
UNetFormer	0.688±0.200	0.796±0.159
UNetFormer2 (Proposed)	**0.718±0.183**	**0.822±0.135**

The proposed model demonstrates that the CBAM blocks added to the UNetFormer enhance its ability to detect and localize regions related to wheat yellow-rust disease, significantly improving IoU and F_1 score metrics in RGB, NDVI, and NIR images compared to the experimental baseline. Results highlight how the enhanced model better identifies and emphasizes the relevant semantic regions associated with the disease.

Figure 9 shows some example qualitative results. The predictions in the last column, generated by the proposed UNetFormer2 model, demonstrate several qualitative improvements over the other SOTA models. For instance, in the first and third rows, the UNetFormer2 predictions show larger light green areas closely matching the ground truth, indicating higher accuracy in detecting yellow-rust-affected pixels. Additionally, the boundaries of the diseased regions are more precisely delineated in the UNetFormer2 predictions, as seen in the second and fifth rows, suggesting that the proposed model captures the disease boundaries more accurately. Moreover, the reduced number of missed predictions across various samples highlights the UNetFormer2's ability to minimize missed detections, providing a more comprehensive segmentation of affected areas.

These improvements make the proposed model a robust choice for accurately identifying wheat yellow-rust disease in NIR images.

Finally, Table 4 shows the computational complexity comparisons of the UNetFormer2 with SOTA models regarding giga floating point operations per second (GFLOPs) and inference time in frames per second (FPS). The utilized GPU is NVIDIA Quadro RTX 5000 for all inference time measurements with 224 × 224 × 3 image resolution. UNetFormer2 has a slightly worse FPS value (103.61) than the baseline UNetFormer (104.92), yet it maintains roughly similar GFLOPs, suggesting a more accurate architecture with high efficiency.

Table 4. Computational complexity comparisons.

Architectures	GFLOPs	FPS
U-Net	190.07	20.48
SegNet	245.80	18.44
DLinkNet	51.43	47.90
NestedU-Net	849.3	5.58
DeepLabV3	133.7	18.30
BiSeNet	34.82	80.60
DFANet	2.73	48.12
UNetFormer	17.95	104.92
UNetFormer2 (Proposed)	18. 32	103.61

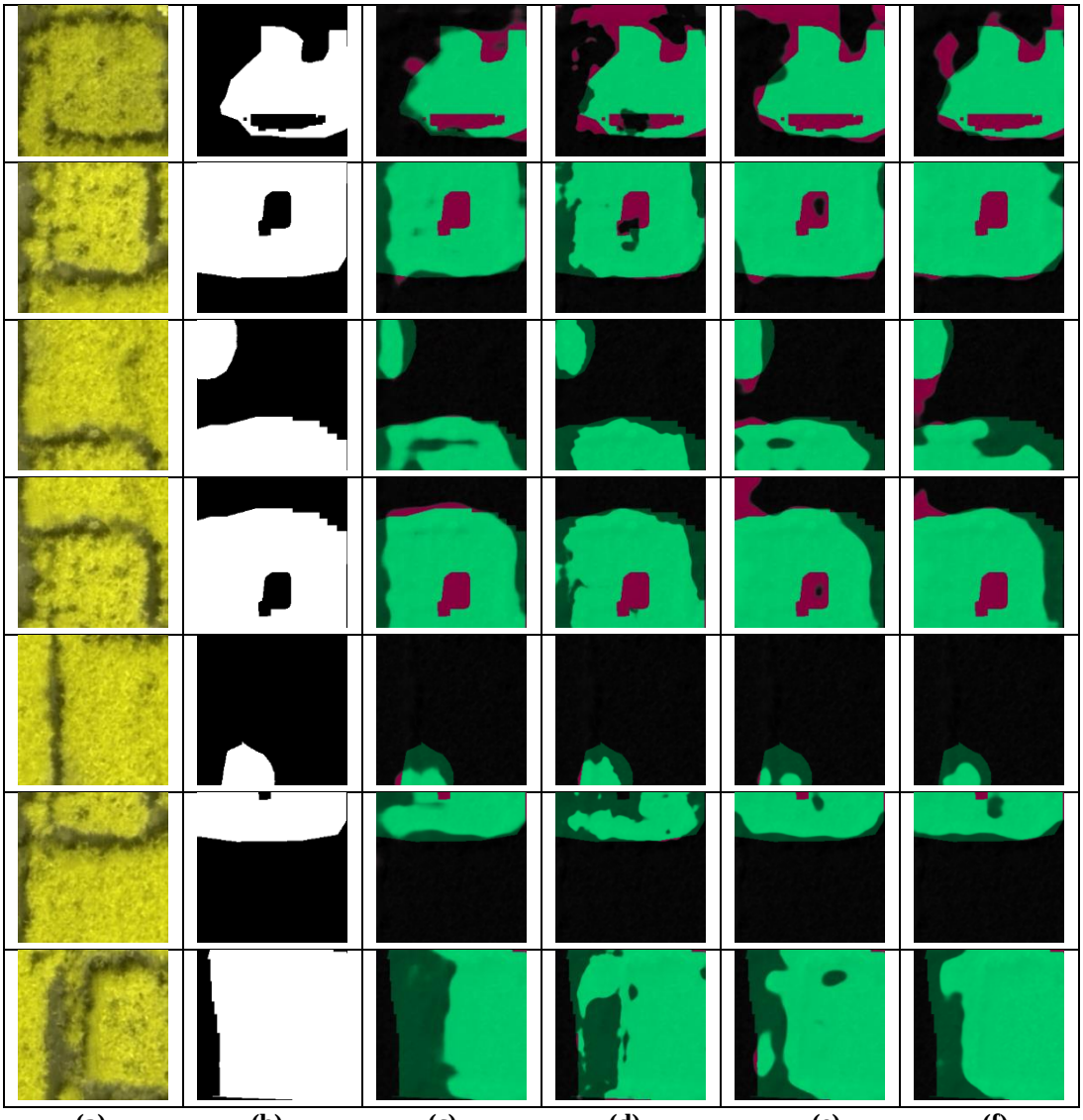


Figure 9. Visual results on NIR test set. Green represents accurate predictions, dark green marks undetected regions, and red highlights false positives. (a) Input images; (b) Ground-truths; (c) BiSeNet; (d) NestedU-Net; (e) UNetFormer; and (f) UNetFormer2 (Proposed).

V. CONCLUSION

This paper proposes UNetFormer2 for wheat yellow-rust disease semantic segmentation in multispectral remote sensing imagery. This approach alleviates the limitations of existing CNN-based models that primarily process local pixel information. UNetFormer2 integrates the CBAM modules into the baseline UNetFormer model to prioritize NIR reflections and boost segmentation accuracy. The CBAM modules improve the discrimination of the relevant features along spatial and channel axes associated with the early disease indicators. These modules focus more on the discriminative NIR reflections to improve the early detection of wheat yellow-rust disease across large agricultural areas.

The experimental results confirm that the UNetFormer2 model has superior performance, underscoring its capability to prioritize crucial features and suppress irrelevant information. Specifically, the model's improved IoU and F_1 scores indicate enhanced precision in detecting disease-affected regions, achieving an IoU improvement of 2.1% for RGB, 4.6% for NDVI, and 3% for NIR compared to baseline. These enhancements facilitate a more reliable solution for wheat yellow-rust disease monitoring, potentially reducing excessive pesticide application and promoting sustainable agriculture by targeting treatment areas accurately.

REFERENCES

[1] Zhang, X., Han, L., Dong, Y., Shi, Y., Huang, W., Han, L., González-Moreno, P., Ma, H., Ye, H., & Sobeih, T. (2019). A deep learning-based approach for automated yellow rust disease detection from high-resolution hyperspectral UAV images. *Remote Sensing*, 11(13), 1554.

[2] Mi, Z., Zhang, X., Su, J., Han, D., & Su, B. (2020). Wheat stripe rust grading by deep learning with attention mechanism and images from mobile devices. *Frontiers in Plant Science*, 11, 558126.

[3] Zhang, J., Pu, R., Loraamm, R. W., Yang, G., & Wang, J. (2014). Comparison between wavelet spectral features and conventional spectral features in detecting yellow rust for winter wheat. *Computers and Electronics in Agriculture*, 100, 79–87.

[4] Liu, W., Yang, G., Xu, F., Qiao, H., Fan, J., Song, Y., & Zhou, Y. (2018). Comparisons of detection of wheat stripe rust using hyperspectral and UAV aerial photography. *Acta Phytopathologica Sinica*, 48(2), 223–227.

[5] Su, J., Yi, D., Coombes, M., Liu, C., Zhai, X., McDonald-Maier, K., & Chen, W. H. (2022). Spectral analysis and mapping of blackgrass weed by leveraging machine learning and UAV multispectral imagery. *Computers and Electronics in Agriculture*, 192, 106621.

[6] Zhang, T., Xu, Z., Su, J., Yang, Z., Liu, C., Chen, W. H., & Li, J. (2021). IR-UNet: Irregular segmentation U-shape network for wheat yellow rust detection by UAV multispectral imagery. *Remote Sensing*, 13(19), 3892.

[7] Su, J., Yi, D., Su, B., Mi, Z., Liu, C., Hu, X., Xu, X., Guo, L., & Chen, W.-H. (2021). Aerial visual perception in smart farming: Field study of wheat yellow rust monitoring. IEEE Transactions on Industrial Informatics, 17(3), 2242–2249.

[8] Ulku, I. (2024). ResLMFFNet: A real-time semantic segmentation network for precision agriculture. *Journal of Real-Time Image Processing*, 21(4), 101.

[9] Su, J., Liu, C., & Chen, W. H. (2022). UAV multispectral remote sensing for yellow rust mapping: Opportunities and challenges. In *Unmanned Aerial Systems in Precision Agriculture: Technological Progresses and Applications* (pp. 107–122). Springer, Singapore.

[10] Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., & Atkinson, P. M. (2022). UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190, 196–214.

[11] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 3–19).

[12] Ulku, I., Tanriover, O. O., & Akagündüz, E. (2024). LoRA-NIR: Low-Rank Adaptation of Vision Transformers for Remote Sensing with Near-Infrared Imagery. *IEEE Geoscience and Remote Sensing Letters*, 21, 1–5.

[13] Ulku, I., Akagündüz, E., & Ghamisi, P. (2022). Deep semantic segmentation of trees using multispectral images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, 7589–7604.

[14] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015* (pp. 234–241).

[15] Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12), 2481–2495.

[16] Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 801–818).

[17] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., & Sang, N. (2018). BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 325–341).

[18] Li, H., Xiong, P., Fan, H., & Sun, J. (2019). DFANet: Deep feature aggregation for real-time semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 9522–9531).

[19] Zhou, L., Zhang, C., & Wu, M. (2018). D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high-resolution satellite imagery road extraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, (pp. 182–186).

[20] Ding, L., Tang, H., & Bruzzone, L. (2020). LANet: Local attention embedding to improve the semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 59(1), 426–435.

[21] Liu, R., Mi, L., & Chen, Z. (2020). AFNet: Adaptive fusion network for remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 59(9), 7871–7886.

[22] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. https://arxiv.org/abs/2010.11929.

[23] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 248–255).

[24] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. *International Conference on Machine Learning*, (pp. 10347–10357).

[25] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (pp. 10012–10022).

[26] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34, 12077–12090.

[27] Strudel, R., Garcia, R., Laptev, I., & Schmid, C. (2021). Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (pp. 7262–7272).

[28] Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H. S., & Zhang, L. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 6881–6890).

[29] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., & Zhou, Y. (2021). TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*. https://arxiv.org/abs/2102.04306.

[30] Yan, Y., Liu, R., Chen, H., Zhang, L., & Zhang, Q. (2023). CCT-Unet: A U-shaped network based on convolution coupled transformer for segmentation of peripheral and transition zones in prostate MRI. *IEEE Journal of Biomedical and Health Informatics*, 27(9), 4341–4351.

[31] Xiao, T., Liu, Y., Huang, Y., Li, M., & Yang, G. (2023). Enhancing multiscale representations with transformer for remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–16.

[32] Lu, C., Zhang, X., Du, K., Xu, H., & Liu, G. (2024). CTCFNet: CNN-Transformer complementary and fusion network for high-resolution remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–17.

[33] Zhang, C., Jiang, W., Zhang, Y., Wang, W., Zhao, Q., & Wang, C. (2022). Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–20.

[34] Wu, H., Huang, P., Zhang, M., Tang, W., & Yu, X. (2023). CMTFNet: CNN and multiscale transformer fusion network for remote-sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–12.

[35] He, X., Zhou, Y., Zhao, J., Zhang, D., Yao, R., & Xue, Y. (2022). Swin transformer embedding UNet for remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–15.

[36] Fan, L., Zhou, Y., Liu, H., Li, Y., & Cao, D. (2023). Combining Swin Transformer with UNet for remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–11.

[37] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 770–778).

[38] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). UNet++: A nested U-Net architecture for medical image segmentation. In *Proceedings of DLMA*, (pp. 3–11).