SCAG-Enhanced U-Net for Wheat Yellow-Rust Semantic Segmentation in Multispectral Remote Sensing

Araştırma Makalesi/Research Article

İrem ÜLKÜ[1*]

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

irem.ulku@ankara.edu.tr

(Geliş/Received:28.02.2025; Kabul/Accepted:16.04.2025)
DOI: 10.17671/gazibtd.1648997

***Abstract*—** The wheat yellow-rust disease poses a serious risk to global wheat production, making effective detection methods essential. This study aims to enhance wheat yellow-rust detection accuracy by investigating the use of spatial-channel attention gates (scAGs) in semantic segmentation with multispectral remote sensing images. While scAGs find applications in medical image segmentation and precision agriculture, this study extends usage for wheat yellow rust detection. Integrated into the skip connections of the U-Net model, scAGs aim to refine feature extraction and improve segmentation performance. Furthermore, to address a limitation in prior work that used only one upsampling method, this study explores multiple techniques—bilinear, bicubic, nearest neighbor, and transposed convolution—optimizing performance. According to experimental results, bicubic interpolation delivers the best performance, significantly enhancing wheat yellow-rust disease detection accuracy.

***Keywords*—** semantic segmentation, upsampling, spatial-channel attention, wheat yellow-rust

SCAG ile Geliştirilmiş U-Net Kullanılarak Çok Bantlı Uzaktan Algılamada Buğday Sarı Pas Hastalığının Semantik Bölütlenmesi

***Özet*—** Buğday sarı pas hastalığı, küresel buğday üretimi için ciddi bir tehdit oluşturmaktadır ve etkili tespit yöntemleri büyük önem taşımaktadır. Bu çalışma, çok bantlı uzaktan algılama görüntülerinde semantik bölütleme için mekânsal-kanal dikkat kapıları (SCAG'ler) kullanımını araştırarak buğday sarı pas hastalığının tespit doğruluğunu artırmayı amaçlamaktadır. SCAG'ler tıbbi görüntü bölütleme ve hassas tarım alanında kullanılmakla birlikte, bu çalışma buğday sarı pas tespiti için kullanımını genişletmektedir. U-Net modelinin atlama bağlantılarına entegre edilen SCAG'ler, özellik çıkarımını iyileştirmeyi ve bölütleme performansını artırmayı hedeflemektedir. Ayrıca, önceki çalışmalar yalnızca tek bir yukarı örnekleme yöntemi kullanırken, bu çalışmada bilineer, bikübik, en yakın komşu ve transpoz konvolüsyon gibi birden fazla teknik araştırılarak performans optimize edilmiştir. Deneysel sonuçlara göre, bikübik enterpolasyon en iyi performansı göstererek buğday sarı pas hastalığının tespit doğruluğunu önemli ölçüde artırmıştır.

***Anahtar Kelimeler*—** semantik bölütleme, yukarı örnekleme, mekânsal-kanal dikkat, buğday sarı pas

1. INTRODUCTION

Modern semantic segmentation frameworks predominantly employ an encoder-decoder design. In this setup, the encoder is responsible for capturing and encoding critical features from input images, while the decoder reconstructs and refines semantic representations to achieve precise segmentation outcomes. However, convolutional neural networks (CNNs) struggle to capture long-distance dependencies inherent in images. Therefore, one study proposes the pyramid pooling module (PPM) to establish and fuse long-distance dependencies [1]. Meanwhile, atrous spatial pyramid pooling (ASPP) is introduced to enhance contextual information by varying the dilation ratios, thereby expanding the receptive field of convolutions [2]. However, these approaches remain inadequate for effectively extracting global contextual information, as stacking and aggregating convolutional layers fail to cover global receptive fields sufficiently.

The attention mechanism is a key for utilizing contextual information. There are several ways to apply attention, where one approach is that the features are aggregated according to the particular attention dimension, followed by linear transformations and nonlinear activations to derive attention scores. For instance, SEBlock employs average pooling and a multilayer perceptron for channel attention [3], and the convolutional block attention module (CBAM) enhances SEBlock by incorporating max pooling for spatial attention [4]. ST-UNet further advances this by aggregating information in both height and width dimensions through soft pooling [5]. Another approach is similarity-based attention, which constructs relationships between units through matrix multiplication. Several studies, such as DANet [6] or MANet [7], fall into this category.

Recent research has shown that attention mechanisms can significantly enhance the performance of feature differentiation and region delineation in remote sensing image analysis. Notably, [8] jointly modeled spatial and channel affinities in remote sensing image segmentation, improving accuracy by addressing attention bias that hinders the discrimination of discretely distributed objects. [9] proposed MCCANet, a multiscale channel-wise cross-attention network incorporating boundary supervision, specifically developed to tackle challenges in segmenting high-spatial-resolution remote sensing images. This approach enhances the integration of multiscale features while maintaining precise boundary delineation. [10] proposed a multistage feature fusion lightweight model, LiANet, integrating enhanced spatial and channel attention modules to significantly reduce parameter count while improving semantic segmentation accuracy for high-resolution remote sensing images. [11] developed SFCRNet, leveraging contextual and multiscale information to enhance the semantic segmentation of remote sensing images using a context-based attention embedding module and a stair-shaped architecture for effective feature fusion. The HMANet [12] improves semantic segmentation in high-resolution aerial images by effectively capturing global dependencies through a novel attention framework that integrates spatial, channel, and category-based correlations. [13] introduced LANet, which enhances feature representation and spatial localization in remote sensing image semantic segmentation by integrating the patch attention and attention embedding modules.

The skip connections in encoder-decoder architectures like U-Net often rely on redundant low-level features and lack sufficient contextual information [14]. By integrating a spatial-channel attention gate (scAG), which incorporates both spatial and channel attention mechanisms, the U-Net model can better emphasize contextual information in the feature maps [15]. This novel approach, designed for medical image segmentation, can also be well-suited to remote sensing semantic segmentation. For instance, the csAG-HRNet model integrates HRNet-v2 with channel and spatial attention gates to effectively learn contextual information for building extraction from aerial images [16]. The GRSNet model enhances U-Net with attention gates, residual units, and deep supervision to effectively segment buildings in high-resolution satellite images [17].

Wheat is one of the most widely grown crops worldwide. However, pathogens and pests, e.g., the wheat yellow rust fungal disease, challenge wheat production. Infections of this disease appear as visible symptoms on wheat leaves parallel to the veins. This disease can reduce wheat production and threaten global food security. Wheat yellow rust spreads rapidly under optimal conditions, necessitating an automated and non-destructive mapping system for effective site-specific management [18]. The disease induces both physical and chemical alterations in wheat leaves, such as a decline in chlorophyll levels and the emergence of rust-like symptoms. Traditional surveillance is the current practice of disease monitoring, which may not be efficient for large-scale fields. Pesticide application is the disease management strategy that eventually results in land pollution. Therefore, it is necessary to develop intelligent disease monitoring methods for early diagnosis.

The wheat yellow rust disease induces both physical and chemical alterations in wheat leaves, such as a decline in chlorophyll levels and the emergence of rust-like symptoms. These changes can be captured through optical sensing technologies, including RGB, multispectral, and hyperspectral imaging systems. Previous studies have utilized UAV-based multispectral remote sensing for mapping yellow rust, employing deep learning models like U-Net. A study investigated a five-band remote sensing-based multispectral camera for detecting yellow rust disease in winter wheat, aiming to distinguish healthy wheat from infected plants by utilizing spectral bands and vegetation indices [19]. The dynamic tracking of wheat affected by varying degrees of yellow rust infection is conducted through time-series aerial imaging with a multispectral camera. This approach enables the assessment of diverse spectral indices, aiding in the

precise segmentation of wheat [20]. A novel framework is proposed for monitoring yellow rust disease in winter wheat using remote sensing-based multispectral imaging, integrating vegetation indices and U-Net [21]. The Ir-UNet model enhances wheat yellow rust detection in UAV-derived multispectral images by integrating non-uniform encoder and decoder components, along with a content-aware channel re-weighting mechanism. This architecture significantly boosts segmentation accuracy, surpassing U-Net [22]. The ResLMFFNet model improves gradient propagation within SEM-B blocks by leveraging residual connections, representing the first application of a real-time semantic segmentation framework for detecting wheat yellow rust disease in multispectral aerial observations [23]. To further improve wheat yellow rust disease detection, attention mechanisms, such as scAG, should be investigated more comprehensively to enhance semantic segmentation accuracy.

Therefore, this study focuses on the semantic segmentation of wheat yellow rust disease using remote sensing-based multispectral images by integrating scAGs into the skip connections of a U-Net-based architecture. While previously used in biomedical imaging tasks [15] and precision agriculture solely for comparison purposes [24], this study represents its first expanded implementation for detecting wheat yellow rust disease. Moreover, the original approach only utilized one upsampling method, leaving the potential benefits of alternative methods unexplored. This study addresses this by adopting bilinear, bicubic, nearest neighbor, and transposed convolution upsampling techniques to optimize performance and identify the best method. The key contributions of this study are as follows:

- Inspired by recent developments in wheat yellow rust disease detection, this study explores the use of attention mechanism of scAGs in the U-Net model design, improving semantic segmentation accuracy on remote sensing-based multispectral data.

- This study also investigates various upsampling techniques, such as bilinear, bicubic, nearest neighbor, and transposed convolution, and shows that bicubic interpolation can increase the performance of scAGAttU-Net model.

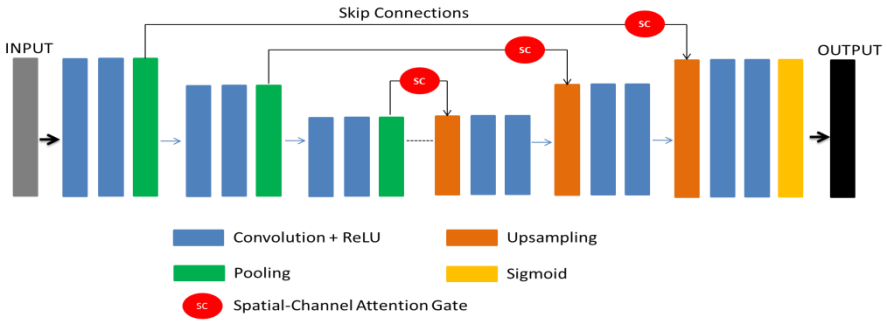


Figure 1. scAGAttU-Net architecture. Input image size is 224 × 224 × 3 pixels. Kernel size of the convolutional layers is 3 × 3 and the max pooling layer size is 2 × 2 with a stride of 2.

2. MATERIALS AND METHODS

Integrating a spatial focus mechanism (sAG) and a channel emphasis module (cAG) into skip connections yields the scAGAttU-Net architecture [15], as depicted in Figure 1.

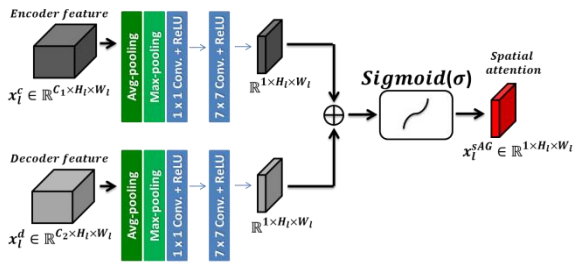


Figure 2. sAG illustration.

Skip connections in the U-Net architecture carry redundant information from low-level features in the encoder at multiple scales. However, using the sAG module (Figure 2), which focuses only on salient spatial regions in low-level feature maps, enables these features to be utilized more efficiently.

The architecture incorporates an extra component, cAG (shown in Figure 3), designed to reduce the semantic disparity between the feature maps generated by the encoder and decoder. The encoder feature map is multiplied with the attention values obtained from the sAG and cAG modules before being combined with the decoder feature map (as illustrated in Figure 4). Equation 1 represents the refined encoder feature map $\hat{x}_l^e$ of the l^{th} layer:

$$\hat{x}_l^{\mathrm{e}} = x_l^{\mathrm{e}} \otimes x_l^{\mathrm{sAG}}\left(x_l^{e}, x_l^{\mathrm{d}}\right) \otimes x_l^{\mathrm{cAG}}\left(x_l^{\mathrm{e}}, x_l^{\mathrm{d}}\right) \quad (1)$$

where $x_l^{\mathrm{e}} \in \mathbb{R}^{\mathrm{C}_1 \times \mathrm{H}_l \times \mathrm{W}_l}$ represents the feature map at the l^{th} layer in the encoder, $x_l^{\mathrm{d}} \in \mathbb{R}^{\mathrm{C}_2 \times \mathrm{H}_l \times \mathrm{W}_l}$ is the feature map of the l^{th} layer in the decoder, $x_l^{\mathrm{sAG}} \in \mathbb{R}^{1 \times \mathrm{H}_l \times \mathrm{W}_l}$ is the spatial attention map of the l^{th} layer, and $x_l^{\mathrm{cAG}} \in \mathbb{R}^{\mathrm{C}_1 \times 1 \times 1}$ is the channel attention map of the l^{th} layer. Activation map height and width of layer l are expressed by H_l and W_l, respectively. C_1 and C_2 denote the number of channels in the feature maps of the encoder and decoder parts, respectively. The symbol $\otimes$ represents element-wise multiplication, which replicates channel-wise emphasis scores across the spatial domain and distributes spatial significance weights across the channel dimension.

By enabling focus on the salient region, the sAG module facilitates the effective utilization of the rich location information in encoder feature maps. Obtaining a spatial attention map involves applying avg pooling, max pooling, and point-wise convolution to the encoder and corresponding decoder feature maps along the channel dimension, as shown in Figure 2. Concatenation occurs between the feature maps $x_l^{\mathrm{sAG,avg}} \in \mathbb{R}^{1 \times \mathrm{H}_l \times \mathrm{W}_l}$, $x_l^{\mathrm{sAG,max}} \in \mathbb{R}^{1 \times \mathrm{H}_l \times \mathrm{W}_l}$, and $x_l^{\mathrm{sAG,1\times1}} \in \mathbb{R}^{1 \times \mathrm{H}_l \times \mathrm{W}_l}$, obtained via avg pooling, max pooling, and point-wise convolution, respectively. The concatenated feature maps undergo convolution with a 7 × 7 kernel size to derive spatial attention maps for the encoder and decoder, denoted as $x_l^{\mathrm{sAG,e}}(x_l^{\mathrm{e}})$ (Equation 2) and $x_l^{\mathrm{sAG,d}}(x_l^{\mathrm{d}})$ (Equation 3), respectively:

$$x_l^{\mathrm{sAG,e}}(x_l^{\mathrm{e}}) = \mathrm{conv}\left(K_1^{7\times7}; \left[x_l^{\mathrm{sAG,avg,e}}, x_l^{\mathrm{sAG,max,e}}, x_l^{\mathrm{sAG,1\times1,e}}\right]\right) \quad (2)$$

$$x_l^{\mathrm{sAG,d}}(x_l^{\mathrm{d}}) = \mathrm{conv}\left(K_1^{7\times7}; \left[x_l^{\mathrm{sAG,avg,d}}, x_l^{\mathrm{sAG,max,d}}, x_l^{\mathrm{sAG,1\times1,d}}\right]\right) \quad (3)$$

where conv(;) stands for the convolution operation, while [ ] denotes concatenation. $K_1^{7\times7}$ corresponds to a single 7 × 7 convolutional kernel. Choosing a large kernel size allows more accurate capture of spatially important regions thanks to the large receptive field. The final spatial attention map, $x_l^{\mathrm{sAG}}\left(x_l^{\mathrm{e}}, x_l^{\mathrm{d}}\right)$, is obtained with the equation (Equation 4):

$$x_l^{\mathrm{sAG}}\left(x_l^{\mathrm{e}}, x_l^{\mathrm{d}}\right) = \sigma\left(x_l^{\mathrm{sAG,e}}(x_l^{\mathrm{e}}) + x_l^{\mathrm{sAG,d}}(x_l^{\mathrm{d}})\right) \quad (4)$$

where σ represents the Sigmoid function.

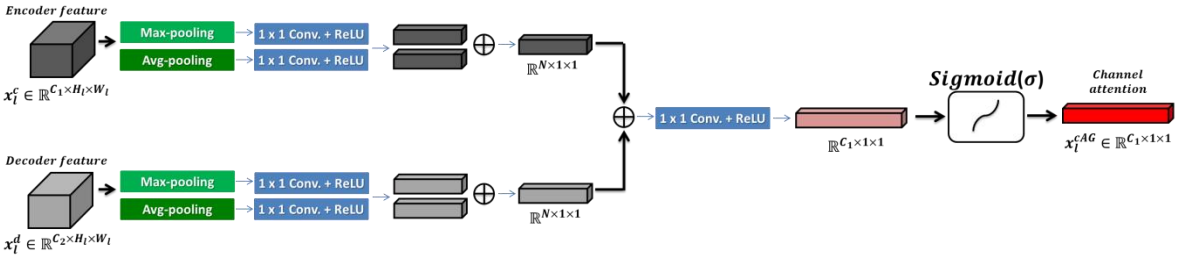


Figure 3. cAG illustration.

Combining low-level and high-level feature maps with skip connections leads to a semantic gap. This gap is mainly due to the low-level features having rich spatial information but inadequate semantic information. The channel attention map is employed to mitigate the semantic gap by enhancing the representational richness of early-stage feature maps. Analyzing inter-channel relationships is essential for assigning appropriate weights, as each channel in the feature map carries distinct information. Assigning higher weights to feature maps containing more discriminant information enhances their effectiveness in the final prediction. Simultaneously utilizing global average pooling and max pooling on the encoder and corresponding decoder feature maps, as illustrated in Figure 3, is essential for generating a channel attention map, as this condenses spatial information. The resulting map of average pooling is denoted as $x_l^{\mathrm{cAG,avg}} \in \mathbb{R}^{\mathrm{C}\times1\times1}$, whereas that of max pooling is $x_l^{\mathrm{cAG,max}} \in \mathbb{R}^{\mathrm{C}\times1\times1}$. To obtain channel-wise dependencies, the squeezed features from the encoder and decoder undergo $C_1/16$ 1 × 1 convolutions. The resulting feature maps are $x_l^{\mathrm{cAG,e}}(x_l^{\mathrm{e}})$ and $x_l^{\mathrm{cAG,d}}(x_l^{\mathrm{d}})$, denoted by Equation 5 and Equation 6, respectively:

$$x_l^{\mathrm{cAG,e}}(x_l^{\mathrm{e}}) = \mathrm{conv}\left(K_{\mathrm{N}}^{1\times1}; \left(x_l^{\mathrm{cAG,avg,e}}\right)\right) + \mathrm{conv}\left(K_{\mathrm{N}}^{1\times1}; \left(x_l^{\mathrm{cAG,max,e}}\right)\right) \quad (5)$$

$$x_l^{\mathrm{cAG,d}}(x_l^{\mathrm{d}}) = \mathrm{conv}\left(K_{\mathrm{N}}^{1\times1}; \left(x_l^{\mathrm{cAG,avg,d}}\right)\right) + \mathrm{conv}\left(K_{\mathrm{N}}^{1\times1}; \left(x_l^{\mathrm{cAG,max,d}}\right)\right) \quad (6)$$

where N is taken as $C_1/16$ for the purpose of reducing the parameter overhead.

Aggregating these feature maps through summation and subsequently applying C_1 point-wise 1 × 1 convolution operations to the resulting map from the element-wise

summation is essential for obtaining the final channel attention map.

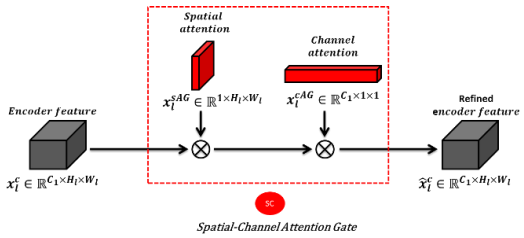


Figure 4. scAG illustration.

Subsequently, the sigmoid function σ is applied, as shown in the Equation 7:

$$x_l^{\mathrm{cAG}}\left(x_l^{\mathrm{e}}, x_l^{\mathrm{d}}\right) = \sigma\left(\mathrm{conv}\left(K_{C_1}^{1\times1}; \left(x_l^{\mathrm{cAG,e}}(x_l^{\mathrm{e}}) + x_l^{\mathrm{cAG,d}}(x_l^{\mathrm{d}})\right)\right)\right). \quad (7)$$

The initial feature representation undergoes multiplication with the location-based attention map $x_l^{\mathrm{sAG}}\left(x_l^{\mathrm{e}}, x_l^{\mathrm{d}}\right)$ and spectral emphasis map $x_l^{\mathrm{cAG}}\left(x_l^{\mathrm{e}}, x_l^{\mathrm{d}}\right)$, as illustrated in Figure 4. Subsequently, these refined features are incorporated with the decoder features, constituting the final step of the process.

2.1. Upsampling Methods

This study examines the effects of bicubic interpolation, nearest interpolation [25], and transposed convolution [26] methods on model performance. Techniques such as nearest neighbor interpolation, along with bilinear and bicubic scaling, belong to a class of resampling methods that fill gaps by leveraging local pixel information. Transposed convolution is a method that learns spatial correlations to extract pixel values. The following sections describe different upsampling methods.

2.2. Nearest neighbor interpolation

This method offers a computationally efficient resampling strategy by assigning the value of the closest pixel to fill missing regions. The interpolation kernel $s(z)$ is defined as follows:

$$s(z) = \begin{cases} 0 & |z| > 0.5 \\ 1 & |z| < 0.5 \end{cases} \quad (8)$$

where z represents the measured distance between the interpolated pixel and the nearest neighbor pixel.

2.3. Bilinear interpolation

This approach fills gaps by calculating the weighted average of the four nearest pixels and is computationally more expensive than the nearest interpolation technique. It also provides smoother transitions between pixels compared to nearest interpolation, as this method performs average calculations in horizontal and vertical directions. The interpolation kernel $s(z)$ is obtained as follows:

$$s(z) = \begin{cases} 0 & |z| > 1 \\ 1 - |z| & |z| < 1 \end{cases} \quad (9)$$

2.4. Bicubic interpolation

This technique fills gaps by computing a weighted combination of the closest 16 pixels. It demands higher computational resources compared to nearest and bilinear interpolation methods. The interpolation kernel is as follows:

$$s(z) = \begin{cases} \frac{3}{2}|z|^3 - \frac{5}{2}|z|^2 + 1 & 0 \le |z| < 1 \\ -\frac{1}{2}|z|^3 + \frac{5}{2}|z|^2 - 4|z| + 2 & 1 \le |z| < 2 \\ 0 & 2 \le |z| \end{cases} \quad (10)$$

2.5. Transposed convolution

Transposed convolution, also referred to as deconvolution, is an upsampling approach with learnable parameters [26]. This process swaps the forward and backward passes of a standard convolution operation. In standard convolution, the forward pass is computed as $C * W$ for a given kernel W and input matrix C, where $*$ represents the convolution operation. The backward pass uses the transposed input matrix C^{T} in the gradient calculation. In contrast, transposed convolution defines the operation of $C^{\mathrm{T}} * W$ for the forward pass, ensuring that during the backward pass, the original input matrix$(C^{\mathrm{T}})^{\mathrm{T}}$, i.e., $(C^{\mathrm{T}})^{\mathrm{T}} = C$, is used.

3. RESULTS AND DISCUSSION

This section presents details on the dataset, assessment metrics, implementation aspects, and experimental findings. The imagery dataset for wheat yellow rust (WYR) analysis was acquired from an agricultural research facility in Yangling, China [21]. The selected wheat cultivar, Xiaoyan 22, was artificially infected with stripe rust spores in randomly designated field plots, each measuring 2 m × 2 m.

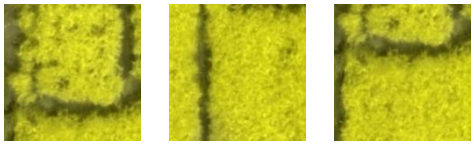
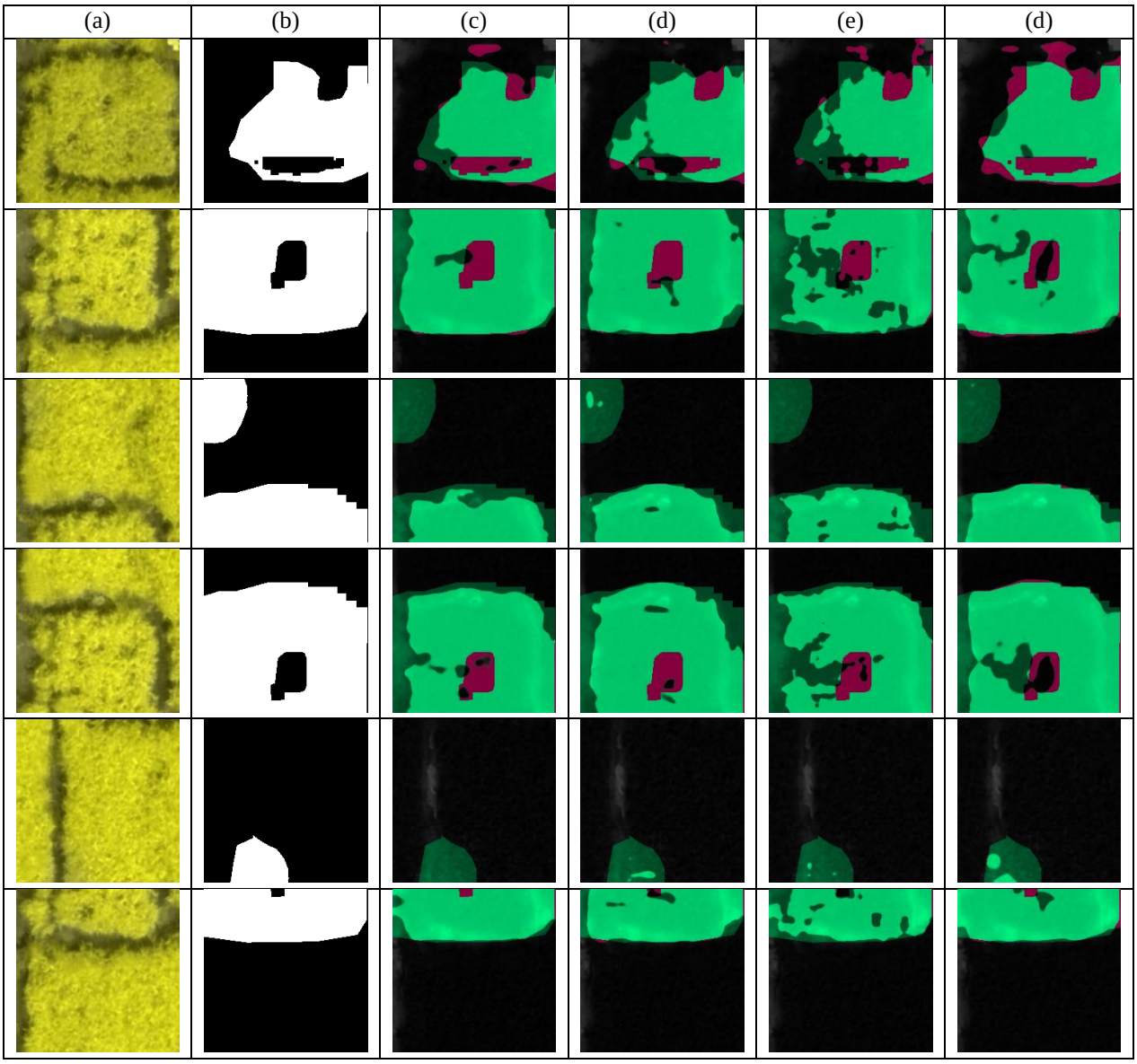


Figure 5. Example images and ground truth masks.

A DJI Matrice 100 (M100) drone system, outfitted with a dual-spectrum optical sensor, is utilized to monitor stripe rust infection. This sensor captures RGB, red-edge (RE), and near-infrared (NIR) imagery at a spatial resolution of 1.3 cm per pixel. The infected regions are annotated at the pixel level and used for binary semantic segmentation. Figure 5 displays sample images alongside their respective ground-truth segmentation maps from the WYR dataset.

The experiments utilize an NVIDIA Quadro RTX 5000 GPU. After preprocessing, all images are divided into 224 × 224 patches, resulting in 1,299 samples. Training samples account for 72% of this total, test samples represent 20%, and validation samples make up 8%. The experiments utilize 5-fold cross-validation, with a total of 70 epochs. The mini-batch size is 8, and the momentum value is 0.9 for the Adam optimization algorithm. The learning rate starts at $5 * 10^{-5}$ and is reduced by 9% every ten epochs.



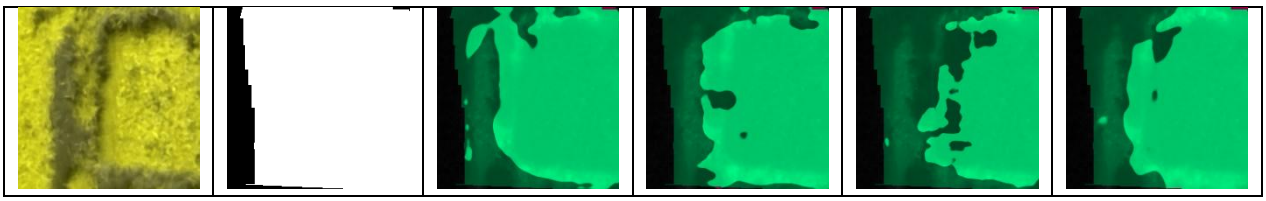


Figure 6. Randomly selected semantic segmentation test results for wheat yellow-rust pixels with RGB images. The input/output image dimensions are 224 × 224 × 3 pixels. In the outputs, light greens indicate correct predictions, dark greens represent missed predictions, and reds denote false alarms. The rows depict the following: (a) Image; (b) Ground-truth mask; (c) scAGAttU-Net (Bilinear); (d) scAGAttU-Net (Nearest neighbor); (e) scAGAttU-Net (Transposed conv.); and (f) scAGAttU-Net (Bicubic).

Table 1. Semantic segmentation test results for RGB images.

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
InceptionU-Net	0.588±0.262	0.699±0.248
NestedU-Net	0.615±0.264	0.718±0.242
UNetFormer	**0.664±0.241**	**0.769±0.203**
DLinkNet	0.608±0.296	0.702±0.295
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
scAGAttU-Net (Bilinear)	0.584±0.285	0.686±0.261
scAGAttU-Net (Nearest neighbor)	0.572±0.296	0.672±0.295
scAGAttU-Net (Transposed conv.)	0.536±0.314	0.632±0.319
scAGAttU-Net (Bicubic)	0.604±0.279	0.707±0.263

Table 2. Semantic segmentation test results for NDVI images.

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
InceptionU-Net	0.569±0.264	0.662±0.252
NestedU-Net	0.560±0.292	0.653±0.279
UNetFormer	0.515±0.320	0.615±0.313
DLinkNet	0.472±0.298	0.575±0.304
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
scAGAttU-Net (Bilinear)	0.596±0.259	0.710±0.331
scAGAttU-Net (Nearest neighbor)	0.573±0.274	0.684±0.254
scAGAttU-Net (Transposed conv.)	0.540±0.289	0.647±0.282
scAGAttU-Net (Bicubic)	**0.618±0.249**	**0.729±0.220**

The normalized difference vegetation index (NDVI) is a popular tool for distinguishing diseased plants [23]. This index can effectively differentiate between healthy and infected wheat leaves by emphasizing the pixels with the highest chlorophyll absorption. NDVI is calculated based on the difference in the ratio of reflectance between the red (R) and NIR bands [27], as shown in Equation 11. From +1 to -1, a positive NDVI value indicates healthy vegetation, while a negative value signifies unhealthy or absent vegetation.

$$\text{NDVI} = (NIR - R)/(\text{NIR}+\text{R}) \tag{11}$$

In this study, NDVI images, which have the potential to be highly useful for distinguishing wheat yellow rust disease, are utilized in addition to RGB images.

The F_1 score and intersection over union (IoU) metrics quantitatively evaluate semantic segmentation performances. Equation 12 shows the calculation of the F_1 score, while Equation 13 yields the IoU metric:

$$\text{F}_1 = (2 * \text{PR} * \text{RC})/(\text{PR} + \text{RC}) \tag{12}$$

$$\text{PR} = TP/(\text{TP} + \text{FP})$$

$$\text{RC} = TP/(\text{TP} + \text{FN})$$

$$\text{IoU} = TP/(\text{TP} + FP + \text{FN}) \tag{13}$$

where TP, TN, FP, and FN correspond to the counts of correctly classified positives, correctly classified negatives, misclassified positives, and misclassified negatives in the prediction results.

In the experiments, deep learning models for semantic segmentation, such as U-Net [14], SegNet [28], InceptionU-Net [29], NestedU-Net [30], UNetFormer [31], DLinkNet [32], DeepLabV3 [33], BiSeNet [34], and DFANet [35] are used for comparison purposes. Table 1 shows the semantic segmentation test results for RGB images based on IoU and F_1 score values, while Table 2 displays the results for NDVI images.

Figure 6 illustrates predictions from scAGAttU-Net (Bilinear), scAGAttU-Net (Nearest neighbor), scAGAttU-Net (Transposed conv.), and scAGAttU-Net (Bicubic) models for selected RGB image examples. The first

column displays the raw image, while the second column presents the corresponding ground-truth segmentation mask. Columns (c), (d), (e), and (f) feature predictions obtained with bilinear, nearest neighbor, transposed convolution, and bicubic upsampling methods, respectively.

Table 3 compares efficiency using giga floating point operations per second (GFLOPs) for computational complexity and frame per second (FPS) for inference speed. For a fair comparison, 224 × 224 × 3 is the input image resolution used in all experiments. The workstation has NVIDIA Quadro RTX 5000 as the GPU.

Table 3. Efficiency comparisons in terms of GFLOPs and FPS metrics.

Architectures	**GFLOPs**	**FPS**
U-Net	190.07	20.48
SegNet	245.80	18.44
InceptionU-Net	482.26	9.48
NestedU-Net	849.3	5.58
UNetFormer	17.95	104.92
DLinkNet	51.43	47.90
DeepLabV3	133.7	18.30
BiSeNet	34.82	80.60
DFANet	2.73	48.12
scAGAttU-Net (Bilinear)	101.51	21.31
scAGAttU-Net (Nearest neighbor)	101.03	22.57
scAGAttU-Net (Transposed conv.)	84.54	24.74
scAGAttU-Net (Bicubic)	113.46	14.60

To explore the differences among the various upsampling methods used in the scAGAttU-Net model with greater precision, Figure 7 and Figure 8 visualize 5-fold cross-validation IoU results for RGB and NDVI images as box plots. Moreover, to compare the statistical characteristics of the scAGAttU-Net variants with other models (U-Net, NestedU-Net, UNetFormer), Figure 9 visualizes the median, interquartile range, and outliers of the 5-fold cross-validation IoU results.

Finally, Figure 10 shows the training and validating loss curves of scAGAttU-Net architectures across different upsampling techniques. These curves can give more intuition about model performances by assessing the stability during training, the convergence behavior, and overfitting.

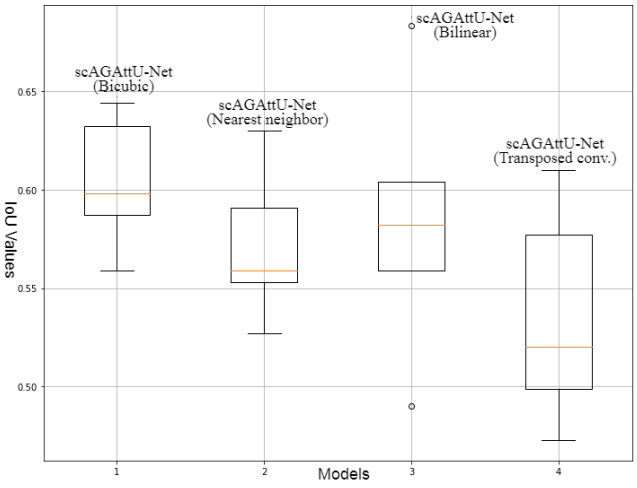


Figure 7. Box-plot of RGB IoU results of different upsampling methods.

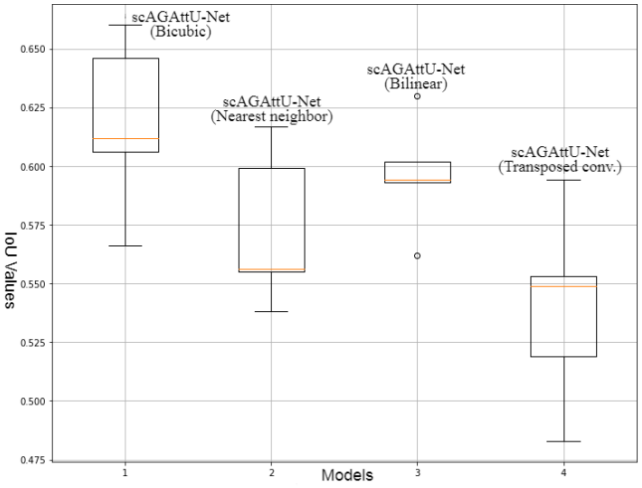


Figure 8. Box-plot of NDVI IoU results of different upsampling methods.

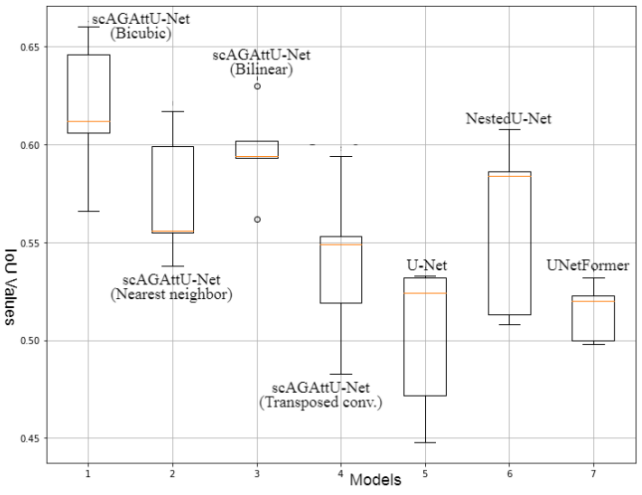


Figure 9. Box-plot of NDVI IoU results of various models.

As indicated in Table 1, UNetFormer, a transformer-based hybrid model, achieves the best performance for RGB images. When examining the scAGAttU-Net model with different upsampling methods, it is evident that bicubic interpolation outperforms the other upsampling techniques. For instance, the scAGAttU-Net (Bicubic) model improves the IoU performance of the scAGAttU-Net (Transposed conv.) model by up to 6.8%.

Table 2 shows that for NDVI images, the scAGAttU-Net (Bicubic) model achieves the highest semantic segmentation accuracy, surpassing all other advanced models significantly. Specifically, the scAGAttU-Net (Bicubic) model improves the IoU performance of the scAGAttU-Net (Transposed conv.) model by up to 7.8%, the scAGAttU-Net (Nearest neighbor) model by 4.5%, and the scAGAttU-Net (Bilinear) model by 2.2%.

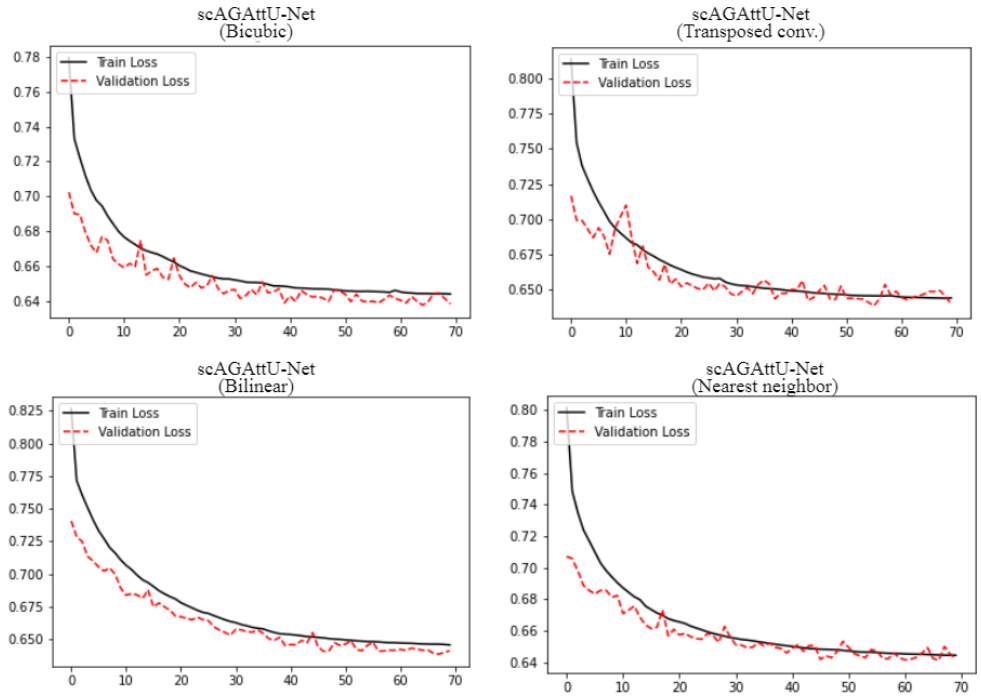


Figure 10. Training and validating loss curves of scAGAttU-Net architectures across different upsampling techniques.

The results from Tables 1 and 2 demonstrate that using bicubic interpolation with scAGAttU-Net yields the best performance across all cases, while bilinear interpolation provides the second-best performance. Moreover, the transposed convolution, which adaptively learns the upsampling process during training, yields the worst performance for RGB and NDVI images on the WYR image set. Within the framework of the scAGAttU-Net model applied to multispectral remote sensing images for wheat yellow-rust disease, bicubic interpolation offers more consistent results than transposed convolution, which may introduce undesirable artifacts despite its ability to learn spatial features.

Figure 6 presents qualitative results from several example predictions on the WYR image set, complementing the quantitative results and allowing for a more detailed analysis of the differences between the upsampling methods. The visualization highlights false alarm pixels in red and shows pixels belonging to correct predictions in light green. Example predictions from the scAGAttU-Net (Bicubic) model shown in the last column demonstrate a decrease in false alarms and an increase in correct predictions. For instance, in the third row, scAGAttU-Net (Bicubic) accurately predicts wheat yellow rust disease pixels. In the fourth row, it is noticeable that using scAGAttU-Net (Bicubic) results in fewer false alarm pixels.

As shown in Table 3, among the scAGAttU-Net model variations, scAGAttU-Net (Transposed conv.) attains superior efficiency with 24.74 FPS and 84.54 GFLOPs. In contrast, scAGAttU-Net (Bicubic) is about 1.69 times slower in inference speed (16.40 FPS) while having 1.34 times more computational cost (113.26 GFLOPs). In Table 1 and Table 2, it is evident that scAGAttU-Net (Bicubic) consistently produces high IoU compared to other scAGAttU-Net models while being the worst performer in terms of efficiency. This fact reveals the efficiency versus accuracy trade-off involved in achieving optimized segmentation performance with various upsampling techniques.

Figure 7 illustrates the distributions of the cross-validation IoU results for RGB images to conduct a comparative analysis among the scAGAttU-Net (Bicubic), scAGAttU-Net (Nearest neighbor), scAGAttU-Net (Bilinear), and scAGAttU-Net (Transposed conv.) models. The box plot indicates that the scAGAttU-Net (Bicubic) model is superior to the others.

Similarly, the distributions presented in Figure 8 for the NDVI images reveal that the scAGAttU-Net (Bicubic) model outperforms the others. The box-plot results further confirm that employing the bicubic interpolation method for upsampling has a statistically significant impact on performance compared to the other methods.

Figure 9 visually demonstrates that the median IoU of the scAGAttU-Net (Bicubic) model is higher than those of U-Net, NestedU-Net, and UNetFormer. Moreover, scAGAttU-Net (Bicubic) has the highest third quartile with a relatively narrow interquartile range, indicating that performance is consistently high across different folds.

In Figure 10, the validation loss of scAGAttU-Net (Bilinear) follows the training loss more closely with minimal fluctuations than other models. Therefore, by allowing more stable training, scAGAttU-Net (Bilinear) has a strong generalization capability while converging relatively faster.

5. CONCLUSION

To summarize, this research effectively showcases the effectiveness of incorporating spatial-channel attention gates (scAGs) into the U-Net model for segmenting wheat yellow-rust disease in multispectral remote sensing imagery. By focusing on different upsampling techniques, the research highlights the superiority of bicubic interpolation, which consistently outperforms other methods, including bilinear, nearest neighbor, and transposed convolution. Specifically, the scAGAttU-Net (Bicubic) model improves the IoU performance by up to 6.8% for RGB images and up to 7.8% for NDVI images compared to the scAGAttU-Net (Transposed conv.) model. Additionally, bicubic interpolation reduces false alarms and increases correct predictions, achieving superior performance across all cases. This study proves that optimizing the scAGAttU-Net model with bicubic interpolation upsampling for wheat yellow-rust disease semantic segmentation using multispectral remote sensing images yields promising results, showing great potential for highly accurate disease detection.

Funding
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest
The author declares that there is no conflict of interest.

Data availability statement
The datasets analyzed during the current study are available from the author on reasonable request.

REFERENCES

[1] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, "Pyramid scene parsing network", **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**, Honolulu, USA, 2881–2890, 2017.

[2] L. C. Chen, G. Papandreou, F. Schroff, H. Adam, "Rethinking atrous convolution for semantic image segmentation", arXiv preprint arXiv:1706.05587, 5, 2017.

[3] J. Hu, L. Shen, G. Sun, "Squeeze-and-excitation networks", **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**, Salt Lake City, USA, 7132–7141, 2018.

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

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

[6] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, "Dual attention network for scene segmentation", **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, Long Beach, USA, 3146–3154, 2019.

[7] R. Li, S. Zheng, C. Zhang, C. Duan, J. Su, L. Wang, P. M. Atkinson, "Multiattention network for semantic segmentation of fine-resolution remote sensing images", IEEE Transactions on Geoscience and Remote Sensing, 60, 1–13, 2021.

[8] X. Li, F. Xu, F. Liu, X. Lyu, Y. Tong, Z. Xu, J. Zhou, "A synergistical attention model for semantic segmentation of remote sensing images", IEEE Transactions on Geoscience and Remote Sensing, 61, 1–16, 2023.

[9] J. Zheng, A. Shao, Y. Yan, J. Wu, M. Zhang, "Remote sensing semantic segmentation via boundary supervision-aided multiscale channelwise cross attention network", IEEE Transactions on Geoscience and Remote Sensing, 61, 1–14, 2023.

[10] R. Guan, M. Wang, L. Bruzzone, H. Zhao, C. Yang, "Lightweight attention network for very high-resolution image semantic segmentation", IEEE Transactions on Geoscience and Remote Sensing, 61, 1–14, 2023.

[11] J. Liu, W. Hua, W. Zhang, F. Liu, L. Xiao, "Stair fusion network with context refined attention for remote sensing image semantic segmentation", IEEE Transactions on Geoscience and Remote Sensing, 62, 1–17, 2024.

[12] R. Niu, X. Sun, Y. Tian, W. Diao, K. Chen, K. Fu, "Hybrid multiple attention network for semantic segmentation in aerial images", IEEE Transactions on Geoscience and Remote Sensing, 60, 1–18, 2021.

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

[14] O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional networks for biomedical image segmentation", **Medical Image Computing and Computer-Assisted Intervention International Conference**, Munich, Germany, 234–241, 2015.

[15] T. L. Khanh, D. P. Dao, N. H. Ho, H. J. Yang, E. T. Baek, G. Lee, S. B. Yoo, "Enhancing U-Net with spatial-channel attention gate for abnormal tissue segmentation in medical imaging", Applied Sciences, 10(17), 5729, 2020.

[16] S. Seong, J. Choi, "Semantic segmentation of urban buildings using a high-resolution network (HRNet) with channel and spatial attention gates", Remote Sensing, 13(16), 3087, 2021.

[17] S. Molavi Vardanjani, A. Fathi, K. Moradkhani, "Grsnet: Gated residual supervision network for pixel-wise building segmentation in remote sensing imagery", *International Journal of Remote Sensing*, 43(13), 4872–4887, 2022.

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

[19] J. Su, C. Liu, M. Coombes, X. Hu, C. Wang, X. Xu, W. H. Chen, "Wheat yellow rust monitoring by learning from multispectral UAV aerial imagery", Computers and Electronics in Agriculture, 155, 157–166, 2018.

[20] J. Su, C. Liu, X. Hu, X. Xu, L. Guo, W. H. Chen, "Spatio-temporal monitoring of wheat yellow rust using UAV multispectral imagery", Computers and Electronics in Agriculture, 167, 105035, 2019.

[21] J. Su, D. Yi, B. Su, Z. Mi, C. Liu, X. Hu, W. H. Chen, "Aerial visual perception in smart farming: Field study of wheat yellow rust monitoring", IEEE Transactions on Industrial Informatics, 17(3), 2242–2249, 2020.

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

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

[24] I. Ulku, "ContexNestedU-Net: Efficient Context-Aware Semantic Segmentation Architecture for Precision Agriculture Applications Based on Multispectral Remote Sensing Imagery", Traitement du Signal, 41(5), 2425-2436, 2024.

[25] E. A. Nogueira, J. P. Felix, A. U. Fonseca, G. Vieira, J. C. Ferreira, D. S. Fernandes, F. Soares, "Upsampling of unmanned aerial vehicle images of sugarcane crop lines with a Real-ESRGAN", **Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering**, Regina, Canada, 285–290, 2023.

[26] M. D. Zeiler, D. Krishnan, G. W. Taylor, R. Fergus, "Deconvolutional networks", **Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition**, San Francisco, USA, 2528–2535, 2010.

[27] A. K. Bhandari, A. Kumar, G. K. Singh, "Feature extraction using normalized difference vegetation index (NDVI): A case study of Jabalpur city", Procedia Technology, 6, 612–621, 2012.

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

[29] I. Delibaşoğlu, M. Çetin, "Improved U-Nets with inception blocks for building detection", *Journal of Applied Remote Sensing*, 14(4), 044512, 2020.

[30] Z. W. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, J. M. Liang, "U-Net++: A nested U-Net architecture for medical image segmentation", **Proceedings of Medical Image Computing and Computer-Assisted Intervention**, Granada, Spain, 3–11, 2018.

[31] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, P. M. Atkinson, "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, 2022.

[32] L. Zhou, C. Zhang, M. Wu, "D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high-resolution satellite imagery road extraction", **Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops**, Salt Lake City, UT, USA, 182–186, 2018.

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

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

[35] H. Li, P. Xiong, H. Fan, J. Sun, "DFANet: Deep feature aggregation for real-time semantic segmentation", **Proceedings of IEEE Conference on Computer Vision and Pattern Recognition**, Long Beach, CA, USA, 9514-9523, 2019.