IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 41, No. 5, October, 2024, pp. 2425-2436
Journal homepage: http://iieta.org/journals/ts

ContexNestedU-Net: Efficient Context-Aware Semantic Segmentation Architecture for Precision Agriculture Applications Based on Multispectral Remote Sensing Imagery

Irem Ulku

Department of Computer Engineering, Ankara University, Ankara 06830, Turkey

Corresponding Author Email: irem.ulku@ankara.edu.tr

https://doi.org/10.18280/ts.410517

Received: 16 December 2023
Revised: 31 March 2024
Accepted: 10 July 2024
Available online: 31 October 2024

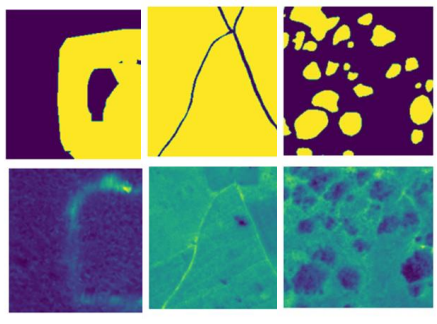
Keywords:
remote sensing, semantic segmentation, precision agriculture

ABSTRACT

Precision agriculture relies on semantic segmentation models to optimize crop yield and minimize environmental impact. ContexNestedU-Net is proposed to improve the capture of contextual information for efficient utilization of multispectral remote sensing images in precision agriculture applications. For this purpose, it includes a novel redesign of the convolutional blocks in the Nested U-Net model. Through the application of depthwise separable convolution in the convolution blocks, the ContexNestedU-Net efficiently preserves unique spectral information. Subsequently, dilated convolution is applied to capture rich contextual information. Three image sets are utilized in the experiments, one from the WorldView-3 satellite and the others from aerial vehicles. Extensive experiments demonstrate that the ContexNestedU-Net outperforms other U-Net-based models for various precision agriculture tasks. When using NDVI images, the proposed architecture improves the Jaccard index by 13% for tree objects, 0.9% for crop objects, and 4.5% for wheat yellow-rust objects compared to Nested U-Net. In addition, the ContexNestedU-Net model reduces the number of trainable parameters from 36.63 to 19 compared to Nested U-Net, and the computational complexity (GLOPs) decreases from 849.3 to 302.4.

1. INTRODUCTION

Precision agriculture is a set of cost-effective technologies to maximize yield and minimize environmental impact [1]. Many agricultural practices are applied for these purposes. Crop detection provides an accurate solution to the discrimination of crop and soil, which is known as one of the most challenging tasks in precision agriculture applications [2]. Tree detection is a remarkably effective tool in reducing greenhouse gas emissions which can minimize environmental impact for the benefit of precision agriculture [3]. Plant disease detection at the early stage prevents outbreaks and yield loss while reducing pesticide usage, in addition to minimizing environmental impact [4]. Figure 1 displays examples of tree, crop, and wheat yellow-rust objects.



(a) (b) (c)

Figure 1. Example samples

The U-Net architecture [5] is widely utilized in precision agriculture applications, specifically with multispectral images [6-8]. Over time, this architecture undergoes enhancements through various mechanisms, leading to the emergence of multiple U-Net architecture variations [9-14].

Some research studies incorporate attention mechanisms into the U-Net architecture to address tasks like forest cover detection [15], tree detection [16], crop type mapping [17], land cover identification [18, 19], and, yellow-rust disease severity level detection [20]. In recent years, a growing trend involves integrating residual connections into the U-Net architecture for remote sensing-based tree detection [21-23]. Additionally, the literature reveals several studies that employ recurrent [24, 25] and residual structures [26, 27], or in combination [28], to enhance U-Net architecture for crop recognition using remote sensing imagery. Lastly, one study introduces Ir-UNet [29], which includes irregular encoder and decoder modules along with content-aware channel re-weighting, to detect wheat yellow-rust disease in aerial remote sensing images. Recent literature also introduces the utilization of transformer-based decoder to design U-Net-like model for urban scene segmentation [30].

Contextual information plays a crucial role in early detection, efficient resource allocation, and precise intervention in precision agriculture applications. When employing the U-Net architecture with multispectral images, a key challenge arises as it struggles to extract adequate contextual information [31]. This study is based on the Nested U-Net architecture, which can capture more contextual

information by employing a hierarchical structure with multiple nested levels [32]. However, the high computational complexity stemming from the nested design makes it less practical for precision agriculture applications.

In this study, the first layer in the convolutional block of the Nested U-Net architecture incorporates depthwise separable convolution [33], chosen for its computational efficiency compared to standard convolution. Multispectral images in precision agriculture convey diverse information about crops and vegetation through different spectral bands. Depthwise separable convolution is suitable for precision agriculture applications relying on multispectral images, as it has the potential to preserve distinct spectral characteristics by applying separate convolutions for each channel [34-38]. Moreover, dilated convolution, also known as atrous convolution, is employed as the second layer to capture additional contextual information [39-41]. Given that some feature relationships in precision agriculture applications span a larger area, this approach is more beneficial, as the model considers both the current pixel and the surrounding area when making predictions.

This study proposes the ContexNestedU-Net architecture, specifically designed for precision agriculture applications using multispectral remote sensing data. It enhances the Nested U-Net by improving convolutional blocks in the following ways:

- The first layer employs depthwise separable convolution to preserve unique spectral features while reducing computational complexity and the number of trainable parameters, making the model highly suitable for multispectral imagery.
- The second layer utilizes dilated convolution to capture broad scene context and enhance contextual awareness.

ContexNestedU-Net efficiently learns spectral features, captures rich contextual information, and reduces computational complexity. Preserving the quality of analysis while efficiently managing computational resources is a crucial consideration in precision agriculture, and ContexNestedU-Net addresses this requirement.

The paper is structured as follows: Section 2 describes the ContexNestedU-Net architecture. Section 3 presents extensive experiments using satellite and aerial multispectral image datasets to compare the semantic segmentation performance of the ContexNestedU-Net architecture with other models, including U-Net, Nested U-Net (UNet++), Attention U-Net (AttU-Net), Recurrent Residual Attention U-Net (R2AttU-Net), Categorical Normalization U-Net (DualNormU-Net), Inception U-Net (InceptionU-Net), UNetFormer, and Spatial-Channel Attention U-Net (scAGAttU-Net). Finally, Section 4 discusses the conclusions.

2. MATERIALS AND METHOD

This section explains the proposed ContexNestedU-Net architecture, image sets and implementation details.

2.1 ContexNestedU-Net

The Nested U-Net architecture [11] is obtained by re-designing the skip connections of U-Net as nested and dense skip pathways that combine the high-resolution feature maps in the encoder with those of their corresponding decoder maps. These nested, dense convolution blocks enhance the semantic similarity between concatenated feature maps, which facilitates capturing high-resolution details. The design of skip connections inspired by DenseNet [42] is the essential feature of the Nested U-Net architecture.

The ContexNestedU-Net architecture (Figures 2 and 3) is built upon the foundation of the nested skip pathway utilized in the Nested U-Net model, aiming to enhance its capacity for capturing contextual information.

Eq. (1) shows the output feature map of layer l, denoted by $x_{l,j}$. Here the j index represents the convolution layer of the dense block utilized along the skip pathway:

$$x_{l,j} = \begin{cases} \mathcal{H}\left(x_{l-1,j}\right) & if\ j = 0 \\ \mathcal{H}\left(\left[\left[x_{l,k}\right]_{k=0}^{j-1}, \cup\left(x_{l+1,j-1}\right)\right]\right) & if\ j > 0 \end{cases} \quad (1)$$

where, the $\mathcal{H}(.)$ function denotes the implementation of the convolution block, whereas the $\cup(.)$ function represents the up-sampling layer. The concatenation operation is represented by [.] symbol.

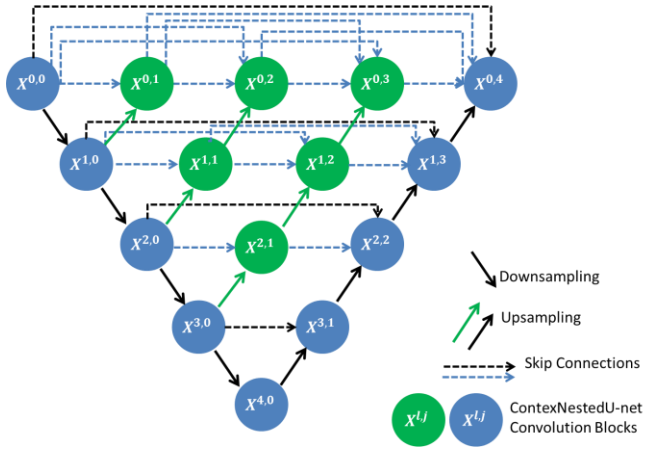


Figure 2. ContexNestedU-Net architecture

As shown in Figure 2, when the convolution layer of the dense block has an index of $j = 0$, it receives input only from the preceding layer in the encoder. However, for an index value of $j \geq 1$, this layer is fed a total of $j + 1$ inputs. These inputs include the outputs of all preceding j layers belonging to the same skipway and the output up-sampled from a sub-skipway. The reason for utilizing all prior feature maps is to incorporate a dense convolution block along each skip pathway. Figure 3 shows the first skip pathway of the ContexNestedU-Net.

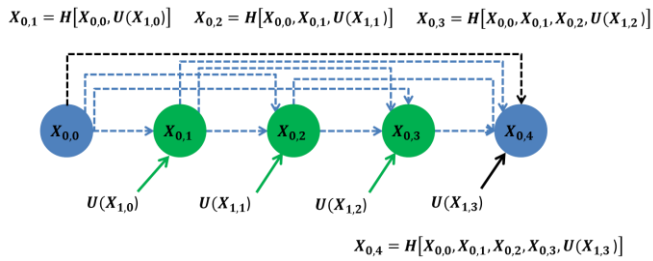


Figure 3. First skip pathway of ContexNestedU-Net

2.1.1 Convolution block

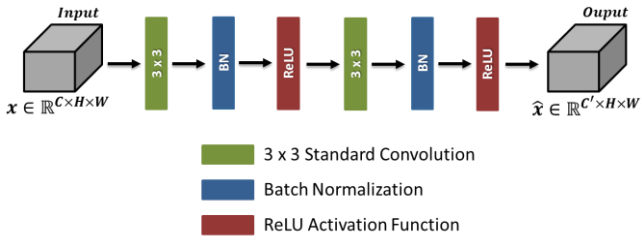
The convolution block of the Nested U-Net architecture

employs two standard convolutional layers with batch normalization and ReLU activation on the input feature map, as visualized in Figure 4 (a).

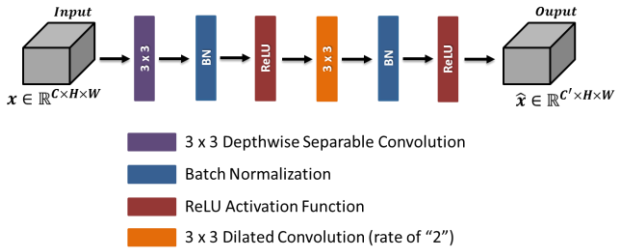
The process of standard convolution, as illustrated in Figure 5 (a), is executed in the following manner, resulting in the expression of the output feature map at layer l+1 as Eq. (2):

$$x_{l+1}^{(c',i,j)} = \sum_{c=1}^{C} \sum_{m=1}^{K} \sum_{n=1}^{K} x_l^{(c,i+m-1,j+n-1)} . \omega^{(c',m,n)} \quad (2)$$

where, C stands for the number of input channels, K represents the size of the convolutional kernel and $x_{l+1}^{(c',i,j)}$ represents the output feature map at layer l+1 with dimensions i (height), j (width) and c' (output channel). x_l is the input feature map at layer l. The term ω refers to the $K \times K$ kernel specific to the output channel c'.



(a) Convolution block of NestedU-Net



(b) Convolution block of ContexNestedU-Net

Figure 4. Convolution block designs

In the ContexNestedU-Net architecture, the objective is to enhance the acquisition of contextual information by redesigning the convolution block. As shown in Figure 5 (b), the new block design replaces the initial standard convolutional layer with a depthwise separable convolutional layer. Since different channels in multispectral image data represent specific characteristics of target objects in precision agriculture applications [43], using depthwise separable convolution is beneficial for capturing the unique spectral features of each channel. Moreover, compared to standard convolution, depthwise separable convolution reduces the number of parameters and computations, which is advantageous for the sustainability of precision agriculture practices.

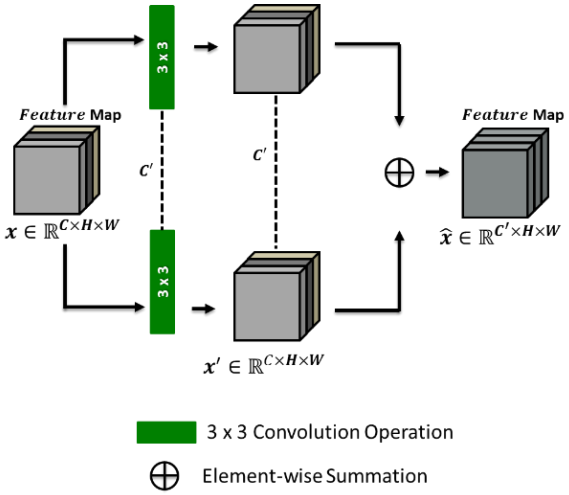
2.1.2 Depthwise separable convolution

As illustrated in Figure 4 (b), the first stage of the two- step depthwise separable convolution process includes the application of depthwise convolution. This operation works independently on each input channel, preserving their distinct characteristics.

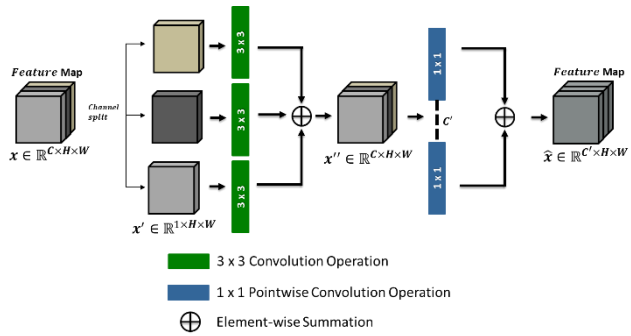
Depthwise convolution generates a set of intermediate feature maps for each input channel l, as follows:

$$x_{l+1}^{(c,i,j)} = \sum_{m=1}^{K} \sum_{n=1}^{K} x_l^{(c,i+m-1,j+n-1)} \omega_d^{(c',m,n)} \quad (3)$$

where, $x_{l+1}^{(c,i,j)}$ denotes the intermediate feature map at layer l+1 acquired for channel c and ω_d represents the kernel associated with channel c.



(a) Standard convolution operation



(b) Depthwise separable convolution operation

Figure 5. Illustrations of convolution operations

As shown in Figure 5 (b), pointwise convolution is applied to merge the intermediate feature maps acquired from all input channels. During this phase, while generating linear combinations of features, the existing unique spectral information remains preserved. This process can even be instrumental in emphasizing significant features while reducing irrelevant information. Therefore, depthwise separable convolution can learn rich features effectively [44].

Pointwise convolution involves the application of a 1 × 1 convolution with c' filters and yields the final output map as outlined below:

$$x_{l+1}^{(c',i,j)} = \sum_{c=1}^{C} x_{l+1}^{(c,i,j)} . \omega_p^{(c',c)}, \quad (4)$$

where, $x_{l+1}^{(c',i,j)}$ denotes the final output feature map belonging to the output channel c' at layer l+1, while $x_{l+1}^{(c,i,j)}$ represents the intermediate feature map at the c channel. For each c output channel, ω_p refers to the 1 × 1 kernel used to combine information from all input channels. Pointwise convolution helps to effectively combine the feature information of

different channels at the same spatial location [45].

ContexNestedU-Net becomes more efficient by employing depthwise separable convolution, which boasts a time complexity of:

$$O \sim (C * K^2 * H * W + C' * C * H * W), \quad (5)$$

Instead of the conventional standard convolution, which carries a time complexity of:

$$O \sim (C' * C * K^2 * H * W), \quad (6)$$

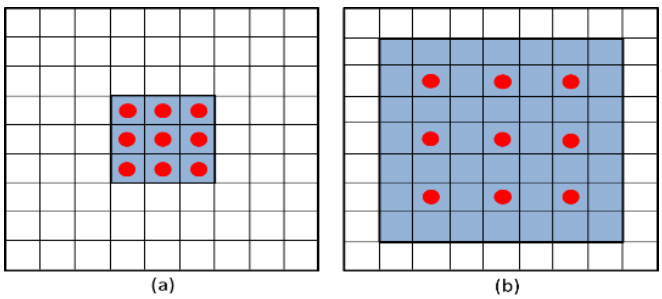
where, C' denotes the output channels, and C represents the input channels, K is the kernel size, while H and W represent the height and width dimensions.

2.1.3 Dilated convolution

In the convolution block of the ContexNestedU-Net architecture, as shown in Figure 4 (b), the second standard convolution operation is replaced with a dilated convolution using a dilation rate of 2. Dilated convolution improves contextual understanding by capturing information from a broader area (refer to Figure 6). Despite this contextual improvement, there is no additional computational overhead. The result is the output feature map at layer $l+1$, which is derived as follows:

$$x_{l+1}^{(c',i,j)} = \sum_{c=1}^{C} \sum_{m=1}^{K} \sum_{n=1}^{K} x_l^{(c,i+(m-1).D,j+(n-1).D)} . \omega^{(c',m,n)} \quad (7)$$

where, $D > 1$ is the dilation rate.



(a) Standard convolution with a kernel size 3 × 3. (b) Dilated convolution with a kernel size 3 × 3 and dilation rate of 2. The blue regions represent receptive fields of the convolutions, and the red circles show the parameters used for the calculations

Figure 6. Dilated convolution operation

Since dilated convolutions allow to extract fine-grained details as well as capture more contextual information, these are used in the convolution block of the ContexNestedU-Net model [46, 47].

2.2 Image sets

In the experiments three remote sensing image sets are used namely DSTL Satellite Imagery Feature Detection Image Set, RIT-18 (The Hamlin State Beach Park) Aerial Image Set and Wheat Yellow-rust Aerial Image Set. The following part describes the details of the image sets.

DSTL satellite imagery feature detection image set: This image set is from the Kaggle competition [48], which contains 25 Worldview-3 [49] satellite images of 1 km × 1 km size, provided by DSTL (Defense Science and Technology Laboratory). The images are labeled pixel-wise for 10 different classes. In this study, only the tree and crop target classes from the DSTL image set are utilized, with each class being considered as a separate binary classification problem. Figure 7 (a) and 7 (b) illustrate an example image and the corresponding ground truth, respectively. The ground truth indicates that the tree class is represented by dark green pixels, while the crop class is denoted by light green pixels.

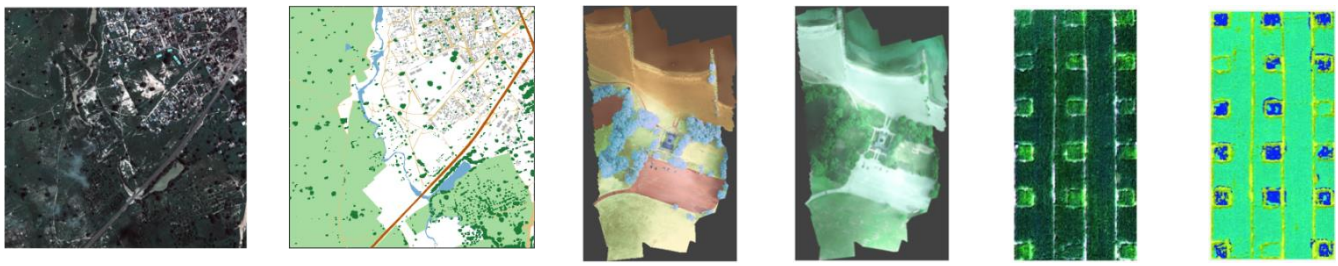
Data is given in the 3-band form consisting of RGB images and in the 16-band form consisting of multispectral images, all of which have different spatial resolutions in GeoTiff format. Panchromatic (P) and RGB images have a size of 3348 × 3392 pixels and the spatial resolution of 0.31 m. Multispectral (M) images have a wavelength range of 400–1040 nm, a size of 837 × 848 pixels, and a spatial resolution of 1.24 m. Short-wave infrared (A) images have a wavelength range of 1195-2365 nm, a size of 134 × 136 pixels, and a spatial resolution of 7.5 m. During the pre-processing step, 5985 image patches are created by resizing all the images into 224 × 224 image patches.

RIT-18 (The Hamlin State Beach Park) aerial image set: Images in RIT-18 [50] were captured from an octocopter with the Tetracam MicroMCA6 multispectral sensor, resulting in a very high-resolution aerial image set. As shown in Figure 7 (c) alongside its corresponding ground truth (Figure 7 (d)), the training image has a size of 9393 × 5642 pixels. Here, the pixels marked in blue represent the tree class.

The RIT-18 image set is known for its high spatial resolution of 0.047 m and contains images in 6 bands, including RGB and near-infrared (NIR), offering exceptional detail. The NIR intervals correspond to the following wavelength ranges: 715-725 nm, 795-805 nm, and 890–910 nm. This study uses the tree class, even though the image set includes 18 distinct pixel-wise labeled class categories. The training image is partitioned into patches of size 224 × 224, resulting in 1778 number samples for training.

Wheat Yellow-rust aerial image set: The collection site is Caoxinzhuang experimental station of Northwest Agriculture and Forestry University, Yangling, China. The DJI Matrice 100 (M100) quadcopter and MicaSense RedEdge multispectral camera captured the images in this aerial image set, which are not publicly available [8]. The seedlings of the Xiaoyan 22 wheat variety were inoculated with a mixture of Pst races (CYR 29/30/31/32/33) [51]. The image in GeoTIFF format at 1336 × 2991 pixel size with a spatial resolution of 0.013 m will be used in this study and shown in Figure 7 (e) with its corresponding ground truth (Figure 7 (f)). After randomly applying yellow-rust inoculum to 2 m × 2 m areas, the affected regions are highlighted in blue in Figure 7 (f) to indicate the presence of the disease.

This aerial image set consists of a total of five bands, i.e., blue (20 nm bandwidth, 475 nm central wavelength), green (20 nm bandwidth, 560 nm central wavelength), red (10 nm bandwidth, 668 nm central wavelength), red edge (10 nm bandwidth, 717 nm central wavelength) and NIR (40 nm bandwidth, 840 nm central wavelength). The division of the image into many 224 × 224 pixels results in a total of 1299 image patches.



(a) (b) (c) (d) (e) (f)

Figure 7. Image set illustrations. (a) An example original image from the DSTL image set. (b) The corresponding ground truth image from the DSTL image set. (c) Original training image from the RIT-18 image set. (d) The corresponding ground truth image from the RIT-18 image set. (e) Original training image from the Wheat Yellow Rust image set. (f) The corresponding ground truth image from the Wheat Yellow Rust image set

2.3 Experimental setup

The implementation of experiments involves the use of the PyTorch framework. The server utilized for training the architectures has the NVIDIA Quadro RTX 5000 GPU. All architectures are trained with adaptive moment estimation (Adam) algorithm [52], where the mini-batch size is 8. During training and validation, the loss function employed is binary cross-entropy with logits. Xavier uniform is used for weight initialization. All image sets undergo training with the architectures for a maximum of 70 epochs. The DSTL and RIT-18 image sets start with an initial learning rate of 10^{-4}, which is decreased by 9% every five iterations. On the other hand, for training the Wheat Yellow-rust image set, the initial learning rate is $5 * 10^{-5}$, which is decreased by 9% every ten iterations. The validation process uses a 5-fold cross-validation approach. The image sets are partitioned into a training set (72%), a test set (20%), and a validation set (8%), with the total number of patches allocated accordingly.

2.4 Evaluation metrics

The Jaccard Index (also known as Intersection over Union or IoU) is the main metric employed in this study to evaluate performance, which is given in Eq. (8) as follows:

$$IoU = TP/(TP + FP + FN) \tag{8}$$

where, *TP* represents pixels that are correctly recognized, while *FP* represents incorrectly identified pixels, and *FN* represents pixels that are not detected. As a second metric to provide a more detailed comparative analysis, the F_1-score calculates the harmonic mean value of precision and recall, which can measure the model's robustness. F_1 -score calculation is given in Eq. (9):

$$F_1\text{-score} = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{9}$$

where, the calculations are $Recall = TP/(TP + FP)$ and $Precision = TP/(TP + FN)$.

3. RESULTS

Table 1 compares the efficiency of the proposed ContexNestedU-Net architecture across various image sets, focusing on memory requirement, computational complexity and inference speed. Computational complexity is measured using model parameters and giga floating-point operations per second (GFLOPs), while frames per second (FPS) measures inference speed.

As shown in Table 1, the ContexNestedU-Net architecture outperforms Nested U-Net in terms of computational cost, reducing GFLOPs by a factor of 2.8. Additionally, ContexNestedU-Net exhibits a remarkable 48.1% reduction in model parameters compared to Nested U-Net. When considering FPS, ContexNestedU-Net is approximately 1.19 times faster than Nested U-Net across all image sets. The memory requirement is reduced by approximately 68.7 MB compared to Nested U-Net.

Table 1. Performance comparison of semantic segmentation architectures estimated on the RIT-18, DSTL, and Wheat Yellow-Rust image sets. Input size is 224 × 224 and GPU is RTX 5000

Architectures	GPU Memory Requirement for Inference (MB)	Params (M)	GFLOPS	RIT-18 Image Set	DSTL Image Set	Wheat Yellow-Rust Image Set
				FPS	FPS	FPS
U-Net	64.09	14.79	190.07	20.54	20.69	20.48
AttU-Net	141.60	34.88	408.12	11.57	11.58	11.55
R2AttU-Net	158.04	39.44	943.42	4.57	4.59	4.58
InceptionU-Net	130.26	32.04	482.26	9.51	9.49	9.43
DualNormU-Net	75.34	17.29	141.13	23.78	23.93	23.87
scAGAttU-Net	40.67	8.66	101.03	22.36	22.60	22.53
UNetFormer	51.95	11.71	17.95	106.66	106.31	104.92
Nested U-Net	147.98	36.63	849.3	5.57	5.60	5.55
ContexNestedU-Net(proposed)	**79.28**	**19.00**	**302.40**	**6.67**	**6.68**	**6.64**

Table 2. Tree semantic segmentation test results in terms of Jaccard Index (IoU) and F_1 score for the different U-Net architectures with DSTL image set

	RGB Images		NDVI Images	
Architectures	**IoU**	***F1***	**IoU**	***F1***
U-Net	0.570 ± 0.206	0.700 ± 0.198	0.462 ± 0.253	0.601 ± 0.252
AttU-Net	0.575 ± 0.206	0.705 ± 0.196	0.487 ± 0.240	0.618 ± 0.231
R2AttU-Net	0.445 ± 0.246	0.570 ± 0.253	0.339 ± 0.275	0.443 ± 0.298
DualNormU-Net	0.548 ± 0.209	0.679 ± 0.211	0.444 ± 0.243	0.574 ± 0.247
InceptionU-Net	0.552 ± 0.203	0.686 ± 0.197	0.486 ± 0.233	0.616 ± 0.240
scAGAttU-Net	0.550 ± 0.216	0.681 ± 0.212	0.477 ± 0.244	0.606 ± 0.245
UNetFormer	0.497 ± 0.213	0.636 ± 0.211	0.390 ± 0.252	0.515 ± 0.263
Nested U-Net	0.549 ± 0.216	0.680 ± 0.200	0.496 ± 0.235	0.627 ± 0.234
ContexNestedU-Net	**0.603 ± 0.412**	**0.712 ± 0.193**	**0.626 ± 0.420**	**0.723 ± 0.237**

Table 3. Crop semantic segmentation test results in terms of Jaccard Index (IoU) and F_1 score for the different U-Net architectures with DSTL image set

	RGB Images		NDVI Images	
Architectures	**IoU**	***F1***	**IoU**	***F1***
U-Net	0.894 ± 0.237	0.904 ± 0.240	0.857 ± 0.285	0.883 ± 0.263
AttU-Net	0.893 ± 0.243	0.902 ± 0.220	0.879 ± 0.275	0.898 ± 0.260
R2AttU-Net	0.857 ± 0.306	0.874 ± 0.296	0.804 ± 0.337	0.836 ± 0.314
DualNormU-Net	0.903 ± 0.234	0.922 ± 0.214	0.882 ± 0.259	0.904 ± 0.238
InceptionU-Net	0.897 ± 0.236	0.919 ± 0.213	0.886 ± 0.252	0.913 ± 0.225
scAGAttU-Net	0.895 ± 0.262	0.910 ± 0.248	0.884 ± 0.279	0.906 ± 0.267
UNetFormer	0.880 ± 0.283	0.896 ± 0.268	0.865 ± 0.307	0.879 ± 0.297
Nested U-Net	0.897 ± 0.234	0.919 ± 0.209	0.880 ± 0.268	0.900 ± 0.251
ContexNestedU-Net	**0.907 ± 0.231**	**0.943 ± 0.180**	**0.889 ± 0.257**	**0.925 ± 0.211**

Table 4. Tree semantic segmentation test results in terms of Jaccard Index (IoU) and F_1 score for the different U-Net architectures with RIT-18 image set

	RGB Images		NDVI Images	
Architectures	**IoU**	***F1***	**IoU**	***F1***
U-Net	0.860 ± 0.285	0.887 ± 0.243	0.841 ± 0.306	0.878 ± 0.269
AttU-Net	0.881 ± 0.265	0.906 ± 0.243	0.883 ± 0.252	0.907 ± 0.227
R2AttU-Net	0.878 ± 0.243	0.904 ± 0.210	0.683 ± 0.464	0.710 ± 0.364
DualNormU-Net	0.880 ± 0.237	0.905 ± 0.215	0.803 ± 0.334	0.828 ± 0.318
InceptionU-Net	0.864 ± 0.269	0.891 ± 0.239	0.873 ± 0.260	0.900 ± 0.236
scAGAttU-Net	0.873 ± 0.271	0.898 ± 0.251	0.860 ± 0.294	0.892 ± 0.262
UNetFormer	0.870 ± 0.260	0.899 ± 0.232	0.842 ± 0.296	0.871 ± 0.272
Nested U-Net	0.885 ± 0.254	0.910 ± 0.228	0.893 ± 0.242	0.918 ± 0.220
ContexNestedU-Net	**0.888 ± 0.251**	**0.914 ± 0.223**	**0.896 ± 0.231**	**0.921 ± 0.205**

Table 5. Wheat yellow rust semantic segmentation test results in terms of Jaccard Index (IoU) and F_1 score for the different U-Net architectures with UAV image set

	RGB Images		NDVI Images	
Architectures	**IoU**	***F1***	**IoU**	***F1***
U-Net	0.521 ± 0.294	0.647 ± 0.333	0.502 ± 0.335	0.582 ± 0.353
AttU-Net	0.553 ± 0.293	0.659 ± 0.289	0.477 ± 0.269	0.529 ± 0.335
R2AttU-Net	0.208 ± 0.310	0.266 ± 0.310	0.292 ± 0.250	0.405 ± 0.257
DualNormU-Net	0.576 ± 0.304	0.672 ± 0.298	0.507 ± 0.324	0.601 ± 0.312
InceptionU-Net	0.588 ± 0.262	0.699 ± 0.248	0.569 ± 0.264	0.662 ± 0.252
scAGAttU-Net	0.563 ± 0.301	0.664 ± 0.298	0.555 ± 0.281	0.644 ± 0.250
UNetFormer	0.664 ± 0.241	0.769 ± 0.203	0.515 ± 0.320	0.615 ± 0.313
Nested U-Net	0.615 ± 0.264	0.718 ± 0.242	0.560 ± 0.292	0.653 ± 0.279
ContexNestedU-Net	**0.691 ± 0.201**	**0.794 ± 0.157**	**0.605 ± 0.258**	**0.717 ± 0.232**

Table 2 displays the test results for semantic segmentation of tree target objects in the DSTL satellite image set. The ContexNestedU-Net architecture outperforms all others for both RGB and NDVI images. In the case of RGB images, it exhibits a 5.4% improvement in the Jaccard index over Nested U-Net. For NDVI images, which incorporate near-infrared data, this improvement reaches a substantial 13%. ContexNestedU-Net excels at preserving discriminative near-infrared reflectance information for vegetation, making it particularly effective for multispectral images.

Table 3 shows test results for semantic segmentation of crop target objects in the DSTL satellite image set. ContexNestedU-Net excels in crop segmentation for RGB and NDVI images, surpassing all other architectures. It enhances Jaccard index by 1% with RGB images and 0.9% with NDVI images when compared to the Nested U-Net architecture. The capacity of

ContexNestedU-Net to improve performance, even with reduced computational costs, stems from its ability to capture rich contextual information.

Similar to the DSTL image set, the RIT-18 aerial image set, which boasts higher spatial resolution, also incorporates the tree target object. Table 4 illustrates test results for tree semantic segmentation within the RIT-18 aerial image set, where ContexNestedU-Net consistently outperforms other architectures. Notably, the proposed ContexNestedU-Net architecture enhances Jaccard index values for RGB and NDVI images by approximately 0.3% compared to NestedU-Net.

Table 5 displays Jaccard index results for yellow-rust disease detection in the Wheat Yellow-Rust aerial image set. Despite the generally low Jaccard index scores due to limited training data, the ContexNestedU-Net model performs exceptionally well by effectively capturing contextual information in all cases. Compared to Nested U-Net, it boosts the Jaccard index by 7.6% for RGB input and 4.5% for NDVI input. The improvement in NDVI performance highlights the model's ability to leverage near-infrared band data effectively for detecting wheat yellow-rust disease. ContexNestedU-Net excels in preserving critical discriminative information, especially related to healthy vegetation found in near-infrared reflectance.

Figure 8 presents a visualization of prediction results using sample images and their corresponding ground truth masks. In the illustrations, light green denotes accurately predicted pixels (hits), dark green represents missed pixels, and red signifies pixels falsely identified (false alarms).

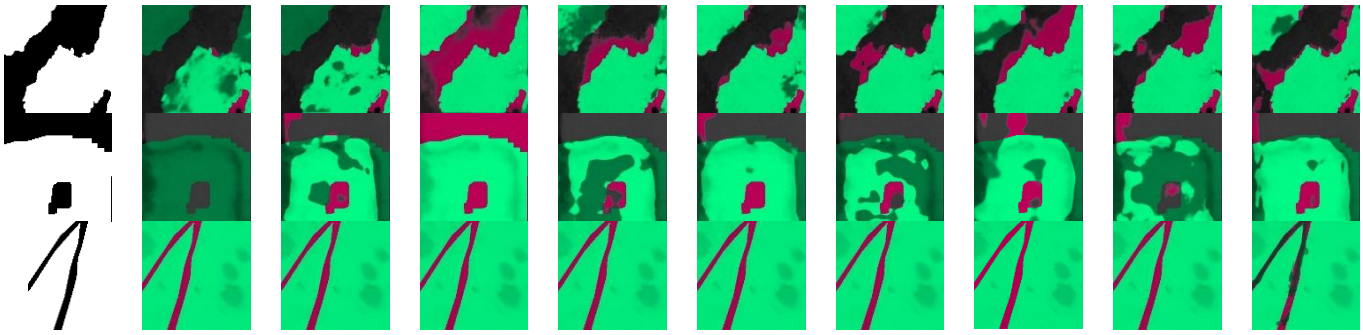
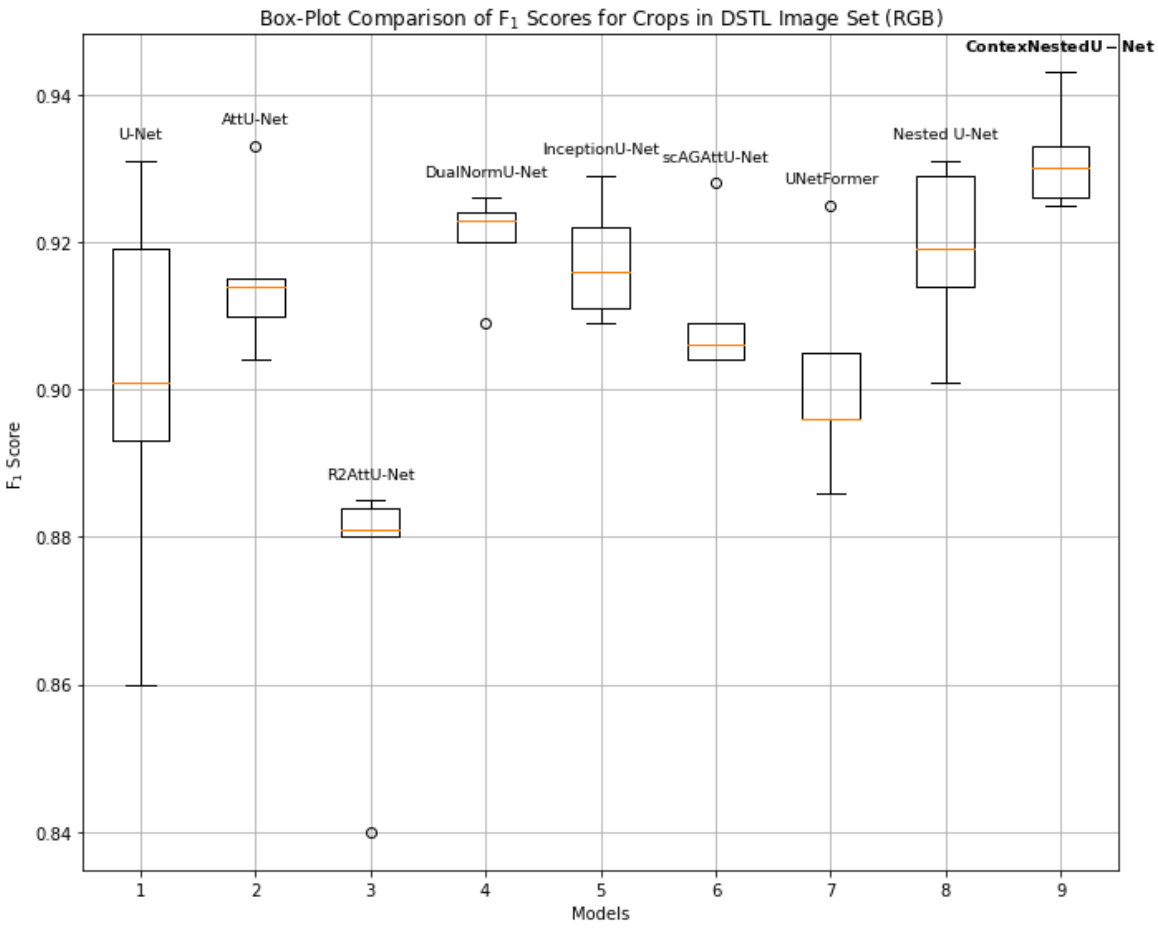
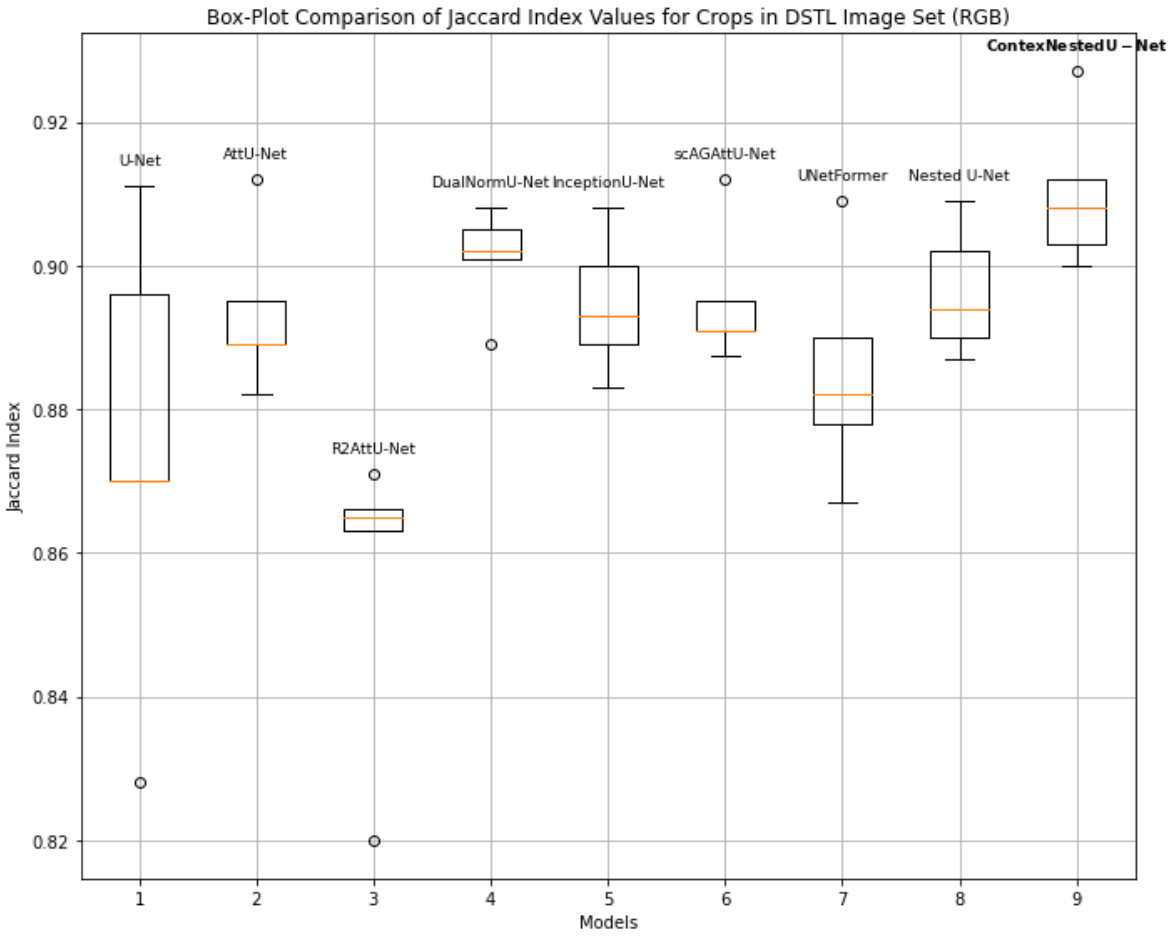


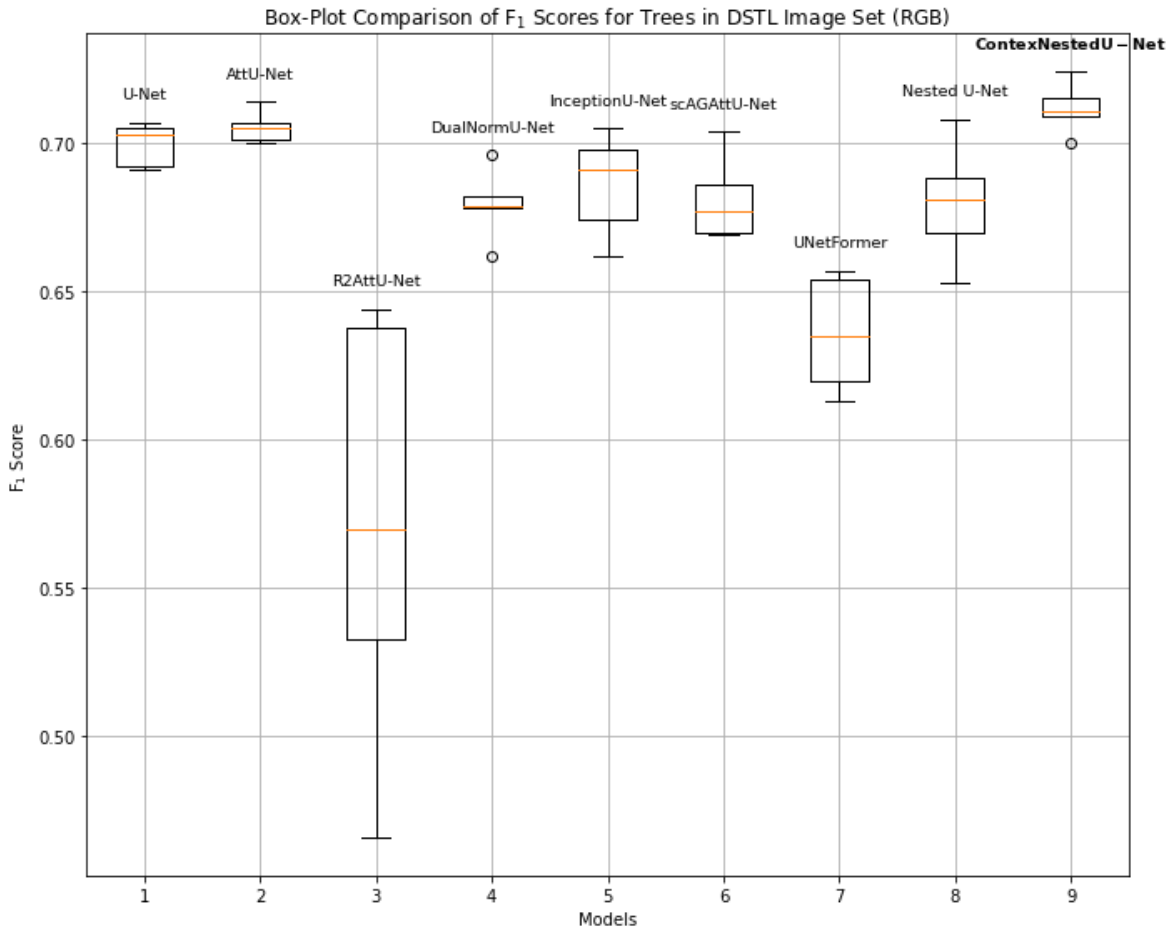
Figure 8. Semantic segmentation test results. Light green represents a hit, dark green represents a miss, and red represents a false alarm. First row shows NDVI tree predictions, second row shows NDVI wheat-yellow rust predictions predictions and third row shows RGB tree predictions. (a) Ground-truth masks. (b) U-Net. (c) AttU-Net. (d) R2AttU-Net. (e) DualNormU-Net. (f) InceptionU-Net. (g) scAGAttU-Net. (h) UNetFormer. (i) Nested U-Net. (j) ContexNestedU-Net



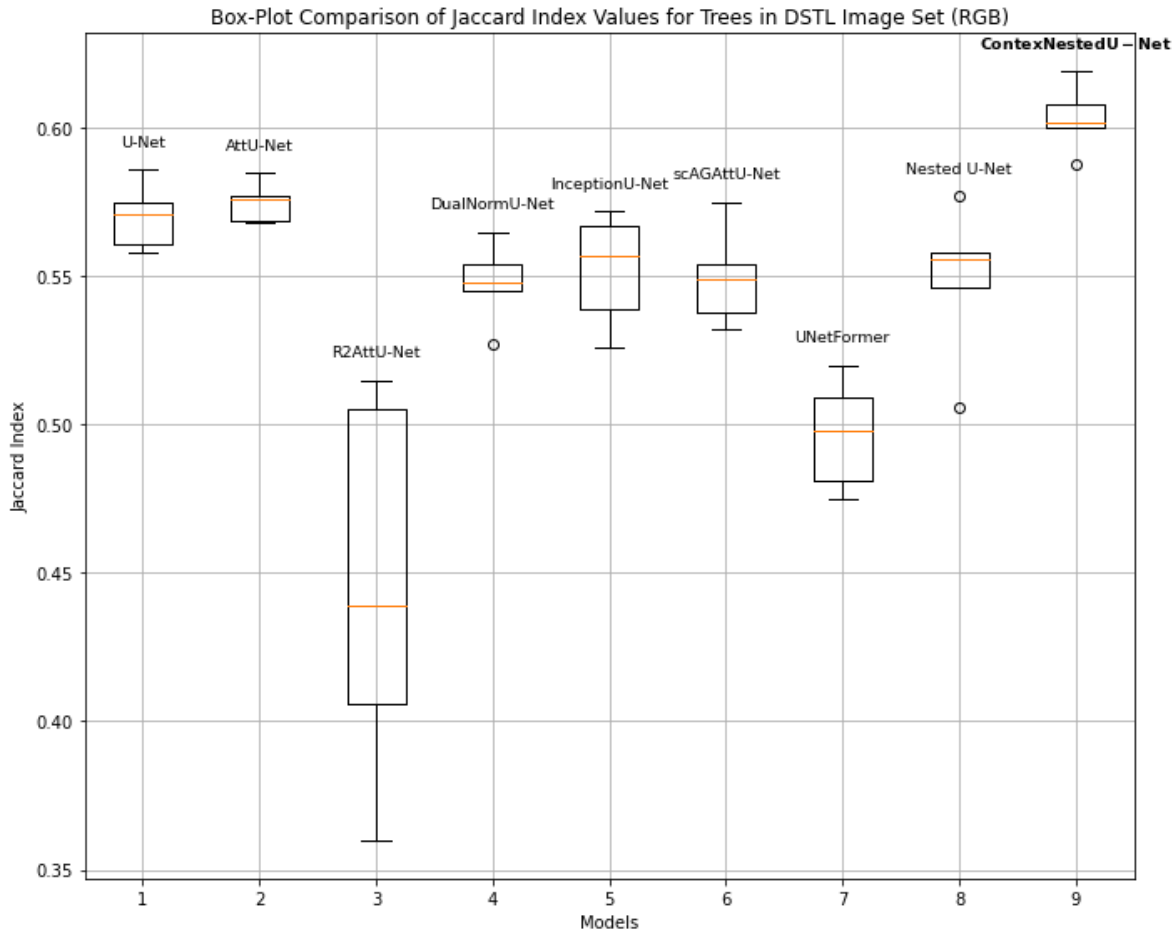
(a)


Box-Plot Comparison of Jaccard Index Values for Crops in DSTL Image Set (RGB)
ContexNestedU – Net
U-Net
AttU-Net
R2AttU-Net
DualNormU-Net
InceptionU-Net
scAGAttU-Net
UNetFormer
Nested U-Net
Jaccard Index
Models
0.92
0.90
0.88
0.86
0.84
0.82
1
2
3
4
5
6
7
8
9

(b)


Box-Plot Comparison of F1 Scores for Trees in DSTL Image Set (RGB)
ContexNestedU – Net
U-Net
AttU-Net
R2AttU-Net
DualNormU-Net
InceptionU-Net
scAGAttU-Net
UNetFormer
Nested U-Net
F1 Score
Models
0.70
0.65
0.60
0.55
0.50
1
2
3
4
5
6
7
8
9

(c)



(d)

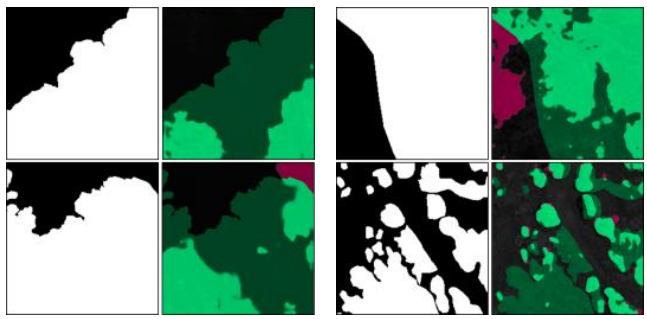
Figure 9. Box-plots of cross-validation results for the different U-Net architectures with DSTL image set. (a) Box-plot comparison of F_1 scores for crops in RGB images. (b) Box-plot comparison of Jaccard Index values for crops in RGB images. (c) Box-plot comparison of F_1 scores for trees in RGB images. (d) Box-plot comparison of Jaccard Index values for trees in RGB images

The first row shows tree predictions using NDVI data from the RIT-18 image set. Notably, the ContexNestedU-Net architecture significantly reduces false alarms; showcasing its ability to effectively preserve the unique spectral characteristics of trees within NDVI input data. In contrast, other architectures tend to misinterpret similar spectral patterns as trees. The second row illustrates the predictions for yellow rust within the NDVI data entries from the Wheat Yellow-rust image set. The ContexNestedU-Net model minimizes both false alarms and missed pixels. The reduction in missed pixels indicates the model's capacity to comprehend the broader contextual information in the scene. The last row shows crop predictions utilizing RGB inputs from the DSTL image set, where the ContexNestedU-Net architecture tends to decrease false alarm pixels.

Figure 9 shows the box-plots over cross-validation results for the DSTL image set. Figure 9 (a) and Figure 9 (b) reveal that ContexNestedU-Net has the highest F_1 scores and Jaccard index values for crop objects, respectively. In Figure 9 (a), the ContexNestedU-Net achieves the highest mean value of 0.931 and the highest median value of 0.930. Similarly, in Figure 9 (b) it obtains the highest mean (0.909) and the highest median (0.907).

Figure 9 (c) and (d) demonstrate that ContexNestedU-Net is also superior in tree objects as compared to other models. It has the highest mean (0.711) and the highest median (0.711) in terms of F_1 score (Figure 9 (c)). It also has the highest median (0.601) and the highest mean (0.603) for Jaccard index values (Figure 9 (d)).

3.1 Limitations



(a) (b)

Figure 10. Prediction masks of ContexNestedU-Net model. (a) Some examples with possible tree occlusion in RIT-18 image set. (b) Some examples with possible tree occlusion in DSTL image set

The proposed ContexNestedU-Net model consistently outperforms other U-Net-based models by capturing more contextual information. However, ContexNestedU-Net still

falls short of predicting possibly occluded tree objects. Some parts of the tree may be blocked by obstacles when there is an occlusion. Examples in Figure 10 (a) and Figure 10 (b) demonstrate that the proposed method is not able to predict possibly occluded trees accurately.

One other challenge is the requirement of having FPS values greater than 30 to be considered as a real-time model. The ContexNestedU-Net is still far from being considered as a real-time model (Table 1).

4. CONCLUSIONS

The ContexNestedU-Net architecture enhances the effectiveness and efficiency of various precision agriculture applications using multispectral remote sensing data. It achieves this by preserving unique spectral information while reducing computational complexity and model parameters through depthwise separable convolution in the first stage of the convolution block. Moreover, the architecture boosts the model's contextual awareness by applying dilated convolution afterwards.

In extensive experiments with diverse multispectral remote sensing image sets, including satellite and aerial imagery, ContexNestedU-Net outperforms various U-Net architectures across all precision agriculture applications. By improving the Nested U-Net's capacity to capture contextual information and maintain unique spectral details, it not only enhances performance but also significantly reduces computational complexity. The proposed ContexNestedU-Net model holds great promise for effectively monitoring large regions, facilitating timely interventions, and enhancing resource utilization in remote sensing-based precision agriculture. As a future work, although the proposed model is faster than the Nested U-Net model, its FPS value needs further improvement to acquire real-time capabilities.

ACKNOWLEDGMENT

The author would like to thank Dr. Jinya Su for sharing the Wheat Yellow-Rust image set [8].

REFERENCES

[1] Micheni, E., Machii, J., Murumba, J. (2022). Internet of things, big data analytics, and deep learning for sustainable precision agriculture. In 2022 IST-Africa Conference, Ireland, pp. 1-12. https://doi.org/10.23919/IST-Africa56635.2022.9845510

[2] Dyson, J., Mancini, A., Frontoni, E., Zingaretti, P. (2019). Deep learning for soil and crop segmentation from remotely sensed data. Remote Sensing, 11(16): 1859. https://doi.org/10.3390/rs11161859

[3] Waters, E., Oghaz, M.M., Saheer, L.B. (2021). Urban tree species classification using aerial imagery. https://doi.org/10.48550/arXiv.2107.03182

[4] Su, J.Y., Liu, C.J., Coombes, M., Hu, X.P., Wang, C.H., Xu, X.M., Li, Q.D., Guo, L., Chen, W.H. (2018). Wheat yellow rust monitoring by learning from multispectral UAV aerial imagery. Computers and electronics in agriculture, 155: 157-166. https://doi.org/10.1016/j.compag.2018.10.017

[5] Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[6] Korznikov, K.A., Kislov, D.E., Altman, J., Doležal, J., Vozmishcheva, A.S., Krestov, P.V. (2021). Using U-Net-like deep convolutional neural networks for precise tree recognition in very high resolution RGB (red, green, blue) satellite images. Forests, 12(1): 66. https://doi.org/10.3390/f12010066

[7] Yang, Q., She, B., Huang, L.S., Yang, Y.Y., Zhang, G., Zhang, M., Hong, Q., Zhang, D.Y. (2022). Extraction of soybean planting area based on feature fusion technology of multi-source low altitude unmanned aerial vehicle images. Ecological Informatics, 70: 101715. https://doi.org/10.1016/j.ecoinf.2022.101715

[8] Su, J.Y., Yi, D.W., Su, B.F., Mi, Z.W., Liu, C.J., Hu, X.P., Guo, L., Chen, W.H. (2020). Aerial visual perception in smart farming: Field study of wheat yellow rust monitoring. IEEE Transactions on Industrial Informatics, 17(3): 2242-2249. https://doi.org/10.1109/TII.2020.2979237

[9] Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, K., Glocker, B., Rueckert, D. (2018). Attention U-Net: Learning where to look for the pancreas. https://doi.org/10.48550/arXiv.1804.03999

[10] Alom, M.Z., Hasan, M., Yakopcic, C., Taha, T.M., Asari, V.K. (2018). Recurrent residual convolutional neural network based on U-Net (r2U-Net) for medical image segmentation. https://doi.org/10.48550/arXiv.1802.06955

[11] Zhou, Z.W., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.M. (2018). Unet++: A nested U-Net architecture for medical image segmentation. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, pp. 3-11. https://doi.org/10.1007/978-3-030-00889-5_1

[12] Xiao, J.F., Yu, L.Q., Zhou, Z.W., Bai, Y.T, Xing, L., Yuille, A., Zhou, Y.Y. (2022). CateNorm: Categorical normalization for robust medical image segmentation. In MICCAI Workshop on Domain Adaptation and Representation Transfer, pp. 129-146. https://doi.org/10.48550/arXiv.2103.15858

[13] Delibasoglu, I., Cetin, M. (2020). Improved U-Nets with inception blocks for building detection. Journal of Applied Remote Sensing, 14(4): 044512. https://doi.org/10.1117/1.JRS.14.044512

[14] Khanh, T.L.B., Dao, D.P., Ho, N.H., Yang, H.J., Baek, E.T., Lee, G., Kim, S.H., Yoo, S.B. (2020). Enhancing U-Net with spatial-channel attention gate for abnormal tissue segmentation in medical imaging. Applied Sciences, 10(17): 5729. https://doi.org/10.3390/app10175729

[15] Yan, C., Fan, X.S., Fan, J.L., Wang, N.Y. (2022). Improved U-Net remote sensing classification algorithm based on multi-feature fusion perception. Remote Sensing, 14(5): 1118.

https://doi.org/10.3390/rs14051118
[16] Li, R., Zheng, S.Y., Duan, C.X., Su, J.L., Zhang, C. (2021). Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images. IEEE Geoscience and Remote Sensing Letters, 19: 1-5. https://doi.org/10.1109/LGRS.2021.3063381
[17] Bian, Y., Li, L.H., Jing, W.P. (2023). CACPU-Net: Channel attention U-Net constrained by point features for crop type mapping. Frontiers in Plant Science, 13: 1030595. https://doi.org/10.3389/fpls.2022.1030595
[18] Liu, J.H., Wang, H., Zhang, Y., Zhao, X.L., Qu, T.F., Tian, H.Z., Lu, Y.T., Su, J.G., Luo, D.S., Yang, Y.L. (2023). A spatial distribution extraction method for winter wheat based on improved U-Net. Remote Sensing, 15(15): 3711. https://doi.org/10.3390/rs15153711
[19] Ma, X.S., Huang, Z.Y., Zhu, S.Y., Fang, W., Wu, Y.L. (2022). Rice planting area identification based on multi-temporal Sentinel-1 SAR images and an attention U-Net model. Remote Sensing, 14(18): 4573. https://doi.org/10.3390/rs14184573
[20] Zhang, T.X., Yang, Z.F., Xu, Z.Y., Li, J.Y. (2022). Wheat yellow rust severity detection by efficient DF-UNet and UAV multispectral imagery. IEEE Sensors Journal, 22(9): 9057-9068. https://doi.org/10.1109/JSEN.2022.3156097
[21] Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C. (2020). ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing, 162: 94-114. https://doi.org/10.1016/j.isprsjprs.2020.01.013
[22] Ran, S., Ding, J.L., Liu, B.H., Ge, X.Y., Ma, G.L. (2021). Multi-U-Net: Residual module under multisensory field and attention mechanism based optimized U-Net for VHR image semantic segmentation. Sensors, 21(5): 1794. https://doi.org/10.3390/s21051794
[23] Zhang, C., Zhang, L., Zhang, B.Y., Sun, J.Q., Dong, S.K., Wang, X.Y., Li, Y.X., Xu, J., Chu, W.K., Dong, Y.W., Wang, P. (2022). Land cover classification in a mixed forest-grassland ecosystem using LResU-Net and UAV imagery. Journal of Forestry Research, 33: 923-936. https://doi.org/10.1007/s11676-021-01375-z
[24] Hao, S.Y., Wang, W., Salzmann, M. (2020). Geometry-aware deep recurrent neural networks for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 59(3): 2448-2460. https://doi.org/10.1109/TGRS.2020.3005623
[25] Martinez, J.A.C., La Rosa, L.E.C., Feitosa, R.Q., Sanches, I.D.A., Happ, P.N. (2021). Fully convolutional recurrent networks for multidate crop recognition from multitemporal image sequences. ISPRS Journal of Photogrammetry and Remote Sensing, 171: 188-201. https://doi.org/10.1016/j.isprsjprs.2020.11.007
[26] Zhou, G.D., Xu, J.H., Chen, W.T., Li, X.J., Li, J., Wang, L.Z. (2023). Deep feature enhancement method for land cover with irregular and sparse spatial distribution features: A case study on open-pit mining. IEEE Transactions on Geoscience and Remote Sensing, 61: 1-20. https://doi.org/10.1109/TGRS.2023.3241331
[27] Zhang, C., Atkinson, P.M., George, C., Wen, Z.F., Diazgranados, M., Gerard, F. (2020). Identifying and mapping individual plants in a highly diverse high-elevation ecosystem using UAV imagery and deep learning. ISPRS Journal of Photogrammetry and Remote Sensing, 169: 280-291. https://doi.org/10.1016/j.isprsjprs.2020.09.025
[28] Zhang, H.X., Liu, M.X., Wang, Y.J., Shang, J.L., Liu, X.L., Li, B., Song, A.Q., Li, Q.Z. (2021). Automated delineation of agricultural field boundaries from Sentinel-2 images using recurrent residual U-Net. International Journal of Applied Earth Observation and Geoinformation, 105: 102557. https://doi.org/10.1016/j.jag.2021.102557
[29] Zhang, T.X., Xu, Z.Y., Su, J.Y., Yang, Z.F., Liu, C.J., Chen, W.H., Li, J.Y. (2021). Ir-unet: Irregular segmentation u-shape network for wheat yellow rust detection by UAV multispectral imagery. Remote Sensing, 13(19): 3892. https://doi.org/10.3390/rs13193892
[30] Wang, L.B., Li, R., Zhang, C., Fang, S.H., Duan, C.X., Meng, X.L., 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. https://doi.org/10.1016/j.isprsjprs.2022.06.008
[31] Soni, A., Koner, R., Villuri, V.G.K. (2020). M-UNet: Modified U-Net segmentation framework with satellite imagery. In Proceedings of the Global AI Congress 2019, pp. 47-59. https://doi.org/10.1007/978-981-15-2188-1_4
[32] Zhou, Q., Wu, X.F., Zhang, S.F., Kang, B., Ge, Z.Y., Latecki, L.J. (2022). Contextual ensemble network for semantic segmentation. Pattern Recognition, 122: 108290. https://doi.org/10.1016/j.patcog.2021.108290
[33] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, pp. 1251-1258. https://doi.org/10.1109/CVPR.2017.195
[34] Weng, L.G., Xu, Y.M., Xia, M., Zhang, Y.H., Liu, J., Xu, Y.Q. (2020). Water areas segmentation from remote sensing images using a separable residual segnet network. ISPRS International Journal of Geo-Information, 9(4): 256. https://doi.org/10.3390/ijgi9040256
[35] Liu, R.C., Jiang, D.W., Zhang, L.L., Zhang, Z.T. (2020). Deep depthwise separable convolutional network for change detection in optical aerial images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13: 1109-1118. https://doi.org/10.1109/JSTARS.2020.2974276
[36] Li, J., Wu, Z.C., Hu, Z.W., Jian, C.L., Luo, S.J., Mou, L.C., Zhu, X.X., Molinier, M. (2021). A lightweight deep learning-based cloud detection method for Sentinel-2A imagery fusing multiscale spectral and spatial features. IEEE Transactions on Geoscience and Remote Sensing, 60: 1-19. https://doi.org/10.1109/TGRS.2021.3069641
[37] Gao, H.M., Yang, Y., Li, C.M., Gao, L.R., Zhang, B. (2020). Multiscale residual network with mixed depthwise convolution for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 59(4): 3396-3408. https://doi.org/10.1109/TGRS.2020.3008286
[38] Shang, R.H., He, J.H., Wang, J.M., Xu, K.M., Jiao, L.C., Stolkin, R. (2020). Dense connection and depthwise separable convolution based CNN for polarimetric SAR image classification. Knowledge-Based Systems, 194: 105542. https://doi.org/10.1016/j.knosys.2020.105542
[39] Chong, F.T., Dong, Z.Y., Yang, X.Z., Zeng, Q.W. (2023). SAR and multispectral image fusion based on dual-

channel hybrid attention block and dilated convolution. In 2023 3rd International Conference on Neural Networks, Information and Communication Engineering (NNICE), Guangzhou, China, pp. 602-607. https://doi.org/10.1109/NNICE58320.2023.10105680

[40] Hu, G.S., Yao, P., Wan, M.Z., Bao, W.X., Zeng, W.H. (2022). Detection and classification of diseased pine trees with different levels of severity from UAV remote sensing images. Ecological Informatics, 72: 101844. https://doi.org/10.1016/j.ecoinf.2022.101844

[41] Zhou, W.J., Fan, X.M., Yu, L., Lei, J.S. (2023). MISNet: Multiscale cross-layer interactive and similarity refinement network for scene parsing of aerial images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 16: 2025-2034. https://doi.org/10.1109/JSTARS.2023.3243247

[42] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, pp. 4700-4708. https://doi.org/10.1109/CVPR.2017.243

[43] Pandya, S., Mistry, M., Parikh, P., Shah, K., Gaharwar, G., Kotecha, K., Sur, A. (2021). Precision agriculture: methodologies, practices and applications. In Proceedings of Second International Conference on Computing, Communications, and Cyber-Security: IC4S 2020, pp. 163-181. https://doi.org/10.1007/978-981-16-0733-2_12

[44] Wang, C.S., Ning, X., Sun, L.J., Zhang, L.P., Li, W.J., Bai, X. (2022). Learning discriminative features by covering local geometric space for point cloud analysis. IEEE Transactions on Geoscience and Remote Sensing, 60: 1-15. https://doi.org/10.1109/TGRS.2022.3170493

[45] Chan, S.X., Wang, Y., Lei, Y.J., Cheng, X., Chen, Z.M., Wu, W. (2023). Asymmetric cascade fusion network for building extraction. IEEE Transactions on Geoscience and Remote Sensing, 6: 2004218-2004218. https://doi.org/10.1109/TGRS.2023.3306018

[46] Safarov, S., Whangbo, T.K. (2021). A-DenseUNet: Adaptive densely connected UNet for polyp segmentation in colonoscopy images with atrous convolution. Sensors, 21(4): 1441. https://doi.org/10.3390/s21041441

[47] Jiang, N., Li, J. (2020). An improved semantic segmentation method for remote sensing images based on neural network. Traitement du Signal, 37(2): 271-278. https://doi.org/10.18280/ts.370213

[48] Kaggle. Dstl Satellite Imagery Feature Detection. https://www.kaggle.com/c/dstl-satellite-imagery-feature-detection, accessed on Dec. 16, 2023.

[49] Satellite Imaging Corporation. WorldView-3 Satellite Imagery. http://www.satimagingcorp.com/satellite-sensors/worldview-3/, accessed on Dec. 16, 2023.

[50] Kemker, R., Salvaggio, C., Kanan, C. (2018). Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. ISPRS Journal of Photogrammetry and Remote Sensing, 145: 60-77. https://doi.org/10.1016/j.isprsjprs.2018.04.014

[51] Su, J.Y., Liu, C., Hu, X.P., Xu, X.M., Guo, L., Chen, W.H. (2019). Spatio-temporal monitoring of wheat yellow rust using UAV multispectral imagery. Computers and Electronics in Agriculture, 167: 105035. https://doi.org/10.1016/j.compag.2019.105035

[52] Kingma, D.P., Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. https://doi.org/10.48550/arXiv.1412.6980