IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 April 2023, accepted 7 May 2023, date of publication 12 May 2023, date of current version 17 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3275748

APPLIED RESEARCH

Learning and Adaptation From Minimum Samples With Heterogeneous Quality: An Investigation of Image Segmentation Networks on Natural Dataset

V. V. SAJITH VARIYAR[1], **V. SOWMYA**[1], **RAMESH SIVANPILLAI**[2], **AND GREGORY K. BROWN**[3]

[1]Centre for Computational Engineering and Networking (CEN), Amrita Vishwa Vidyapeetham, Coimbatore 641112, India
[2]Wyoming GIS Center, School of Computing, University of Wyoming, Laramie, WY 82071, USA
[3]Department of Botany, University of Wyoming, Laramie, WY 82071, USA

Corresponding author: V. Sowmya (v_sowmya@cb.amrita.edu)

This work was supported by Amrita Vishwa Vidyapeetham University, Tamilnadu, India.

ABSTRACT Training deep learning-based image segmentation networks require large number of samples of adequate quality. However, obtaining large number of samples is not possible in certain domains. Recent approaches use augmentation and transfer learning techniques to overcome small sample size. Augmentation techniques are known to introduce noise to the dataset, while transfer learning approaches may fail if the existing dataset is novel to deep learning algorithms. This study investigated how four deep learning-based image segmentation networks learned and adapted to identify epiphytes when trained with fewer image samples (n = 132) of heterogeneous quality without transfer learning and data augmentation. Encoder-Decoder with skip connection (Unet), Deep Residual (DRUnet), Vision transformer (TransUnet), and Conditional Generative (Pix2Pix) represent different generations of deep learning networks. The segmentation performance of the trained models was evaluated by computing the Jaccard score (IoU) for predicted labels for test images. Test images (n = 20) with heterogeneous quality were evaluated by categorizing them into six categories based on target occupancy and lighting conditions. Results from this study showed that among the four networks, predicted images from the TransUnet model achieved high average Jaccard score of 0.78. Role of additional layers apart from Unet was important for accurate localization and context understanding of the target plants. However, these networks misclassified visually similar plants as target plant. The transformer and attention layers in TransUnet showed significant contribution towards improvement in localizing target and understanding context in images with varying quality. TransUnet can be used for segmenting target plants when fewer training samples are available. The presence of Unet based encoder-decoder in TransUnet is well contributing for deriving good features from minimum samples.

INDEX TERMS Image segmentation, less training samples, image quality, encoder-decoder networks, deep neural networks, vision transformer.

I. INTRODUCTION

The deep learning (DL)-based segmentation approaches have proven their capability for various applications in different domains [1], [2]. Segmenting target from an input image requires a good understanding of the context and precise localization of the target [3]. In DL-based solutions, the expressive space of the model must be large enough to adapt well and learn the relevant information.

Data collection in some domains is complex and expensive, making it challenging to build high-quality annotated dataset in sufficient quantity. In domains like environmental science, certain events are observed very rarely. For example, collecting images of rare species/animals resulting in

The associate editor coordinating the review of this manuscript and approving it for publication was Kumaradevan Punithakumar.

fewer samples. This poses challenges for training and testing DL-based networks. The limited sample problem is encountered in domains such as medical imaging and environmental science.

Epiphytes are plants growing in areas like tree trunks and branches. Access to these plants is possible by manually climbing trees or constructing watch towers / temporary structures next to the tree/target. This conventional mode of data collection is risky, time-consuming, and expensive, and therefore results in fewer samples. A previous study reported the use of unmanned aerial vehicles (UAV's) or drones for imaging epiphytes [5]. Drones can reduce the risks during data collection and collect relatively more samples compared to manual methods. However, capturing epiphyte images using drones may compromise the quality of the dataset collected since they are flown inside the canopy. Sajithvariyar et al. reported the challenges and opportunities while flying the UAV inside the canopy are reported in the following work [5]. Due to a host of reasons, only 132 images of heterogeneous quality were suitable for segmentation.

Data augmentation [6] and transfer learning techniques [7] are recommended to overcome the problem of small data samples in deep learning applications. However, geometrical, and color-space transformations add noise and bias to the training images, which may not appear in test images [6]. Further, augmenting techniques require resources such as additional memory, space, and processing time. Transfer learning is a technique to imbibe knowledge from the source domain to a target domain. The knowledge is derived from data trained using DL architectures. It is always recommended to do transfer learning if source and target domain data have similar features; else, the improvements are less significant in models [7]. There are CNN based semantic segmentation networks capable to reduce the utilization of computational resources while training and inference. The semantic segmentation networks named FU-net [27] and REF-net [28] specifically designed to deploy applications in hardware's with limited computational resources. Eventhough these networks are computationally cheap, these networks are trained on large data samples. Apart from the above mentioned networks another network named AEDCN-net [29] claims low computational time and high accuracy. The current study is mainly focused on capability of networks to deals with minimum training samples rather than optimizing computational resources during training and inference.

The epiphyte dataset is novel to the deep learning algorithms. To the best of our knowledge the epiphyte dataset is not introduced to deep learning algorithms for any image processing application. The plant dataset collected using unmanned aerial vehicles or hand held cameras are widely used for detection of diseases, analyzing the plant growth and estimation of yields [30], [31]. The semantic plant segmentation introduced by segmentation introduced by Sakurai et al. used a two-step transfer learning method. A network is initially trained with dataset which contains maximum images from plant domain and later these weights are used while training the dataset with less samples. This study reports that the source and target domain in this study during transfer learning is having very high correlation which helped to retain better weights during transfer learning [32]. The data collection using UAV in agriculture applications for crop growth analysis, yield estimation and disease detection requires the UAV to fly above the plant/crop in an open field. The challenges while flying UAV's inside the canopy are not applicable in open field survey except the lighting conditions [33]. Hence the data collected from these domains will have good quality dataset with reasonable sample size for training a DL model.

Selecting an optimal DL network for segmenting targets with fewer training samples without transfer learning and augmentation requires an effective way of localizing and context understanding while training the deep neural network. Several DL-based networks are available for segmenting targets in images, and comparative analysis to evaluate the performance of few widely used ones is needed.

The fully convolutional neutral (FCN) network proposed by Long et al. [8] localized targets in the images using skip connections that combine lower and coarse layers that take global structure into account. The encoding (or down-sampling) and decoding (or up-sampling) steps in an FCN network localize the target and capture the context respectively. Studies have reported that FCN-based networks with skip connections ensure precise target localization with small number of samples [8]. However, studies have also reported that due to the large size of the filter, some context information is lost during propagation in the decoding steps [8].

Ronneberger et al. introduced Unet for localizing and simultaneously capturing the context from small number of samples [3]. The large number of feature channels between each up-sampling step enables Unet to preserve context. Further, Unet can accurately predict border pixels using weighted loss to distinguish the border pixels. Several studies have reported Unet's ability to predict target class in various 3D medical image segmentation applications [9], [10], [11], [12], [13].

Unet's encoder-decoder combination has been combined with other DL architectures to improve their segmentation ability. Presently many DL architectures have Unet's encoder-decoder for completing their down-scaling and up-scaling data during the training stage [14].

DRUnet has a similar encoder-decoder structure to Unet, with few components derived from RestNet and DenseNet. DRUnet alters the usual Unet encoder path with further concatenation between input and output and decoder blocks with 1×1 convolution to reduce the input dimension. DRUnet is efficient in back propagation of gradients by aggregating feature maps by additional connection between the first and last Conv-BN layers [15].

Recently there are neural networks developed to train in an adversarial fashion. The adversarial trained networks are capable of image generation, which penalizes for an objective function [16]. This competitive mode of training the networks allows better adaptation and learning while building the model. Generative Adversarial Network (GAN) algorithms use generator (Unet) and discriminator networks for adversarial training and are used for various image segmentation tasks [17], [18]. Pix2Pix is a GAN algorithm that is known for its image-to-image translation with fewer training samples [20]. GAN networks used for document image binarization in which pixels in an image are classified as either foreground or background [18]. This process of classification can be challenging when it deals with degradation and noise in the images. Epiphyte segmentation can be challenging with complex background with high similarity between the target (epiphyte) and background (other plants and trees) classes.

The limitations in CNN based neural networks for vision applications [21], [22] are addressed in attention-based transformer networks [23], [24]. TransUnet uses transformer networks with self-attention mechanisms to overcome the limitations while modelling long-range relations among the data [25]. However, the transformers lack target localization due to 1D operations on the sequence of information extracted from the input data. TransUnet feeds the input images to a Unet encoder for extracting features, which are then fed to the transformer blocks as a sequence of embeddings. The output from the transformer blocks are passed to Unet decoder, which upsamples the output from the transformer block and combines the low-level information from the encoder blocks through skip connections. The unique combination of Transformer and Unet decoder makes it a unique network.

The objective of this study is to evaluate the ability of 4 commonly used DL-based networks to learn from relatively fewer samples without data augmentation or transfer learning techniques. In this study, we analyzed the learning and adaptation of four segmentation networks having Unet as the encoder-decoder unit. For this study, we selected three DL networks (DRUnet, Pix2Pix, and TransUnet) that incorporate Unet based encoder-decoder and compared their performance to the original Unet. Our assumption is that presence of Unet and contributions from additional layer in each network will ensure precise localization and context understanding from minimum training samples with heterogeneous quality. All four networks are trained to build an epiphyte segmentation model. The input to the trained model is an epiphyte color image; the output will be a segmented mask with target and background classes separated. All four networks are trained on the same training images and evaluated on the same test images. The learning and adaptation of the four networks are analyzed using the validation loss and accuracy over training iterations. The segmentation performance of each model derived from 4 networks is quantitatively analyzed using the Jaccard Score [26]. The significant contributions and scope of this study are limited to the following.

- Analyzing the learning and adaptation of each network with limited samples for training and heterogeneous quality.

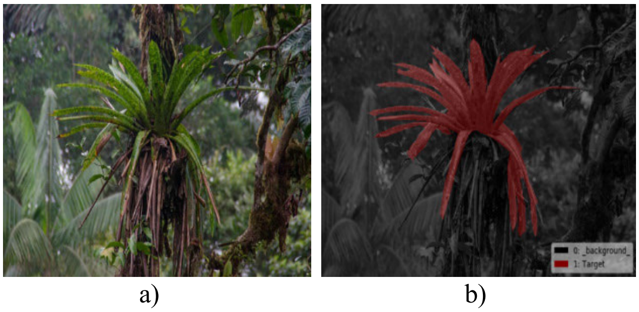

FIGURE 1. a) The epiphyte input image b) The annotated image of epiphyte image where target is highlighted in red pixels and background in black pixels.

- We are exploring the ability of Unet to deal with minimum sample size, precise localization and context understanding for epiphyte data.
- Influence of Unet in DRUnet, Pix2Pix and TransUnet to deal with minimum training samples, precise localization and context understanding for epiphyte data during inference.
- Examine the individual contributions of 4 networks apart from Unet layers, instead of their unique structure and learning mechanism while training with epiphyte data.
- Identify the capability of 4 networks to deal with the heterogeneous quality in test set.

The rest of the study is organized as follows. Section II details the epiphyte dataset used in this study. Section III gives the details of the four networks used for segmentation, and their training strategy and section IV gives the segmentation performance of the model derived from 4 networks and their corresponding quantitative performance comparison. The potential of each model to deal with test images of heterogeneous quality is analyzed by dividing the test images into six categories.

II. THE EPIPHYTE DATASET

This study is performed on an epiphyte dataset which consists of 132 target species images captured using a drone. The input images hold a dimension of 512 × 512 × 3. The current study concentrates on a single epiphyte species named *Warahuai Kupperina* captured from the Costa Rica reserve forest [4], [5]. The quality factors of the epiphyte images, like lighting conditions, the distance at which the drone captures the images, and the occupancy of the target and background, vary throughout the image samples. The various deep neural architectures used in this study follow a supervised learning method that requires the ground truth mask for each image. The epiphyte dataset is pixel-wise annotated into two classes to generate the mask, composed of a target and background, as shown in Fig 1.

A. THE EPIPHYTE DATASET QUALITY

The epiphyte dataset collected is composed of images with varying quality. The epiphytes are plants that grow on top of trees and high altitude places. It will be challenging to collect

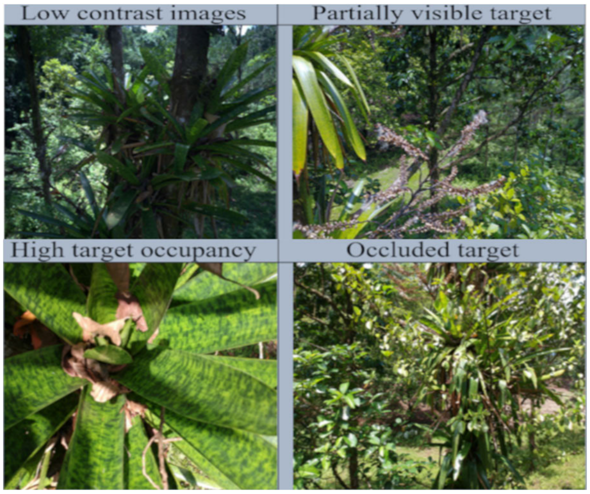

FIGURE 2. The epiphyte images of varying quality used while training and testing the models.

the images with drones uniformly. The dataset is composed of images with varying lighting conditions and distance at which the images is captured as shown in below Fig 2. The deep learning algorithms find it challenging to train a model with a dataset consisting of fewer samples and more variations among the images.

III. DEEP LEARNING MODELS FOR SEGMENTATION

The proposed study implemented the 4 deep learning architectures to train an epiphyte segmentation model. The 4 networks chosen for this study are Unet [3], DRUnet [15], Pix2Pix [20] and TransUnet [25]. The selected deep learning architectures are coming from different generations with unique layers and training strategy. From 132 image samples 107 images are used for training, 5 images for validation and 20 images are used for testing the model. The validation images for TransUnet network is randomly chosen from 112 training samples during training. The test images across the models are made common to compare the performance of each model. The overall training and testing of the 4 models for this task is depicted in Fig 3. The 4 networks shown in Fig 3 is having some similarity by adopting a common encoder-decoder layers and skip connections with in this layers. The encoder-decoder layers present in DRUnet, Pix2Pix and TransUnet follows an original Unet based encoder decoder structure and skip connections. Apart from encoder decoder layers the 4 networks are having additional layers.

The DRUnet follows a Unet structure with additional layers derived from Densenet, Resnet and feature map aggregation blocks which eliminate gradient decent problems while back propagating.

The adversarial mode of training in Pix2Pix generative network makes it different from other networks training strategy. In Pix2Pix the Unet encoder-decoder is only used in generator network to produce fake samples.

Among the 4 networks TransUnet is having a more additional layers compared to other 3 networks. In TransUnet the features from Unet encoder layers are vectorized in to patches of size 16 × 16. These patches are projected into a d dimensional embedding space. From the embedding space the patches undergo a second level training in transformer blocks which contains Multihead Self-Attention (MSA) and Multi-Layer Perceptron (MLP) blocks.

The output from transformer blocks are given to Unet decoder to produce the final output. The Unet, DRUnet, Pix2Pix and TransUnet implemented in this study follows a similar architecture as given in their original implementation. The network parameters of 4 networks while training are given in Table 1.

A. UNET FOR EPIPHYTE SEGMENTATION

Unet is a state of the art segmentation network widely used for medical image segmentation for limited training samples. The unique orchestration of convolutional layers and skip connections during encoding and decoding in Unet makes it different from other networks. The unique property of Unet to deal with minimum samples made it combined with other networks to solve similar problems. Many deep learning models used Unet as the feature extraction network in initial stages to derive robust features for segmentation task. It is worth investigating how epiphyte segmentation is benefitted when Unet is combined with other networks with different training mechanisms.

The proposed study trained an epiphyte segmentation model only with Unet and compared the segmentation scores with the other three methods. We implemented the original Unet implementation proposed by olaf [3] for the epiphyte segmentation task. These inferences will help to understand the opportunities and challenges of epiphyte segmentation techniques used in this study with challenges involved in the dataset and minimum training samples.

B. DRUnet FOR EPIPHYTE SEGMENTATION

It is worth looking of improvements of segmentation when we have additional layers along the Unet layers. The quality of output produced by network depends on the coarser layers which utilize the fundamental features derived in the initial layers. DRUnet is enhanced version of Unet with additional layers. DRUnet is novel with unique capability to produce minimum trainable parameters. The DRUnet is primarily designed for a medical image segmentation application [15]. The DRU-net follows a fundamental Unet architecture with additional Batch normalization operation after every layer. These additional changes help to faster convergence and stabilize network learning.

The DRUnet makes use of the property from densenet, and the residual network makes it ideal for robust image segmentation. The DRUnet network architecture used in this study is depicted in Fig 4. The Unet backbone of this network is well known for training with limited samples and less training time. The epiphyte data set used in this study has challenges like, limited training samples, high pixel overlap between target and background to be segmented, the size and

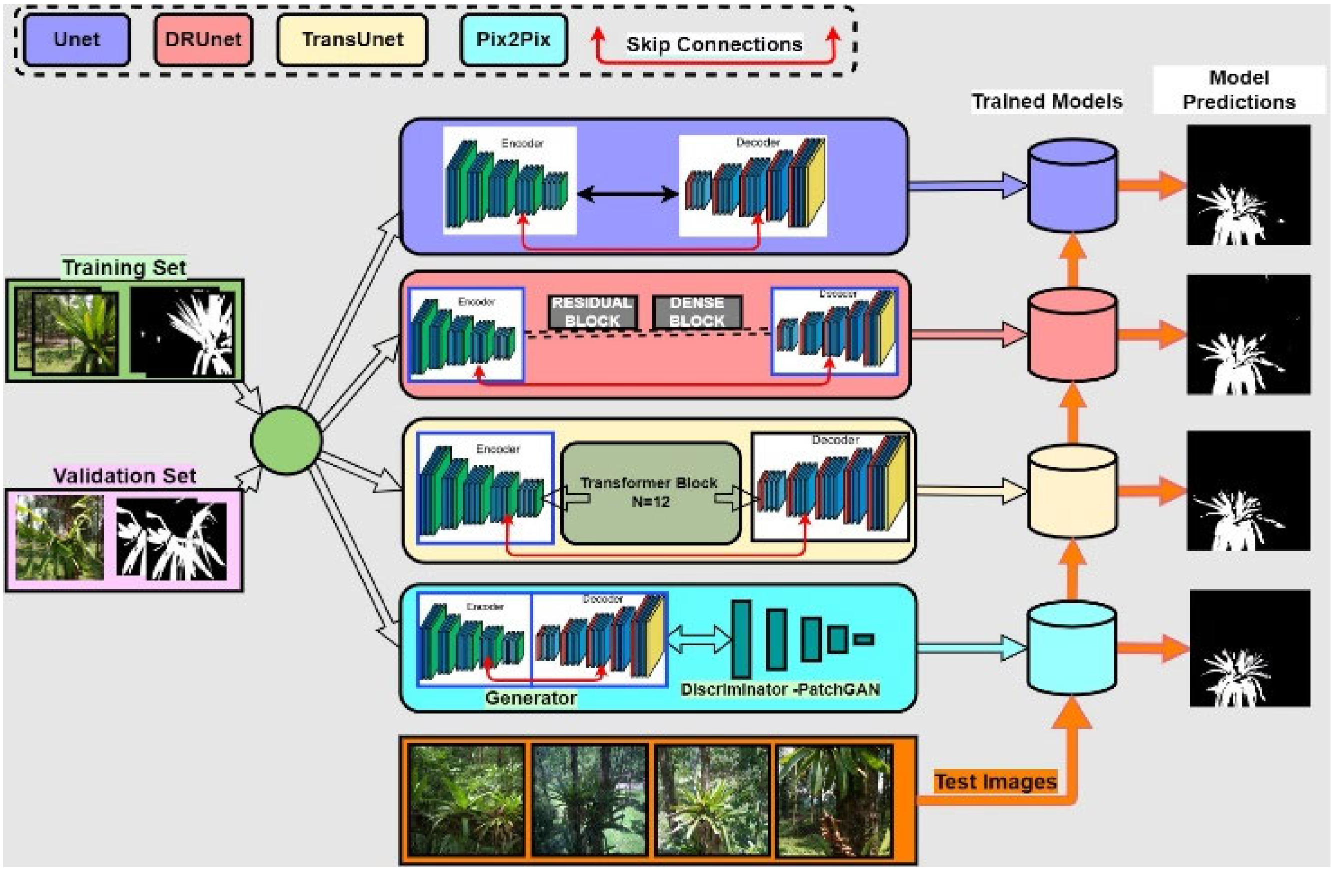

FIGURE 3. The overall training process of the 4 models on epiphyte dataset. All the 4 architectures have common encoder-decoder structure and skip connections.

position of epiphytes in drone images and lighting conditions. The DRUnet derives the property of Resnet and Densenet in the encoder-decoder blocks to get good features with minimum samples and less training overhead. In DRUnet, the additional connections between the first convolution and batch normalization operations to the last convolution –Batch normalization output with a summation operation for feature map aggregation are implemented.

This addition is essential for network training with limited samples like epiphytes to avoid gradient decent problems while back propagating from current to previous convolutional blocks. The 1×1 convolutional blocks are implemented in decoder blocks instead of cropping the output from previous layers reducing information loss in the feature maps. This additional modification in network contributes well while segmenting epiphytes where we have high pixel intensity similarity between target and background.

C. Pix2Pix FOR EPIPHYTE SEGMENTATION

The generative models composed of generator and discriminator network in adversarial settings grabbing high attention in many computer vision application. The generative models are proven for image generation with limited training samples and extensively used as data augmentation technique. In this study we used a variant of GAN named Conditional-GAN primarily introduced in an image to image translation application named Pix2Pix [20]. Unlike the conventional GANS the conditional GAN will keep a copy of the image to be generated as reference to the generator network.

The Pix2Pix network illustration is given in Fig 5. The conditional GAN network used in this study having Unet with skip connections as the generator network and CNN network named patchGAN as the discriminator network. This Conditional GAN with Unet as the image generator makes it ideal to deal with limited training samples. We are also investigating the capability of this generative models to segment epiphyte when we have varying target(s) size /occupancy, lighting conditions and high pixel level similarity between two subjects to be segmented.

The generative models learns the distribution of the data to be generated through an adversarial designed objective functions with two loss functions for generator and discriminator. These networks not only learn the mapping from input image to output image, but also learn a loss function to train this mapping. This will enforce the network to learn data distribution faster compared to other CNN based sequential training techniques for segmentation.

The pix2pix architecture implemented for this study has Unet as the generator network and a convolutional PatchGAN classifier as the discriminator network. The epiphyte dataset

TABLE 1. **The hyper parameters and their settings for the 4 networks used for epiphyte segmentation.**

Parameters	UNet	DRUNet	Pix2Pix	TransUnet
Learning rate	1e-4	Initial value=0.001 later it is adaptive	0.001	2e-4
Batch Size	2	4	2	1
Optimizer	Adam V2	SGD	Adam	Adam
Weight decay	Not set	0.0001	Not set	beta_1=0.5
Optimizer Momentum	Not set	0.9	Not set	0.5
Loss function used	BCE	BCE	BCE	L1 loss for Generator , Discriminator loss

are stored in pairs were each input has its corresponding label image or known as target image. The generator will take the epiphyte color image as the input. During training each target generated from the generator is compared with ground truth target image and output from generator is transferred to discriminator network for computing the loss. The training procedure of the pix2pix network is listed below

- For each input image the generator will generate output image.
- The discriminator receives the input image and the generated image from previous step as the first input.
- The second input to the discriminator is the input image and its ground truth image.
- Then, it will compute the generator loss and the discriminator loss.
- Then, calculate the gradients of loss with respect to both the generator and the discriminator variables (inputs) and apply those to the optimizer.

D. TransUnet FOR EPIPHYTE SEGMENTATION

TransUnet, the visual attention model has proven their capability in medical image segmentation tasks [25]. The CNN based state of the art architecture like Unet is having intrinsic locality of convolution operations, U-Net generally demonstrates limitations in explicitly modelling long-range dependency in the features derived from input images. TransUNet, which merits both Transformers and U-Net, as a strong alternative for medical image segmentation. On the one hand, the Transformer encodes tokenized image patches from a convolution neural network (CNN) feature map as the input sequence for extracting global contexts. On the other hand, the decoder up-samples the encoded features combined with the high-resolution CNN feature maps to enable precise localization. Considering the challenges with the epiphyte dataset, the network should be able to understand the global context and further enable precise localization of the target in the input image. The TransUNet is a hybrid architecture with the UNet encoder connected with Transformer layers as shown in Fig 6.

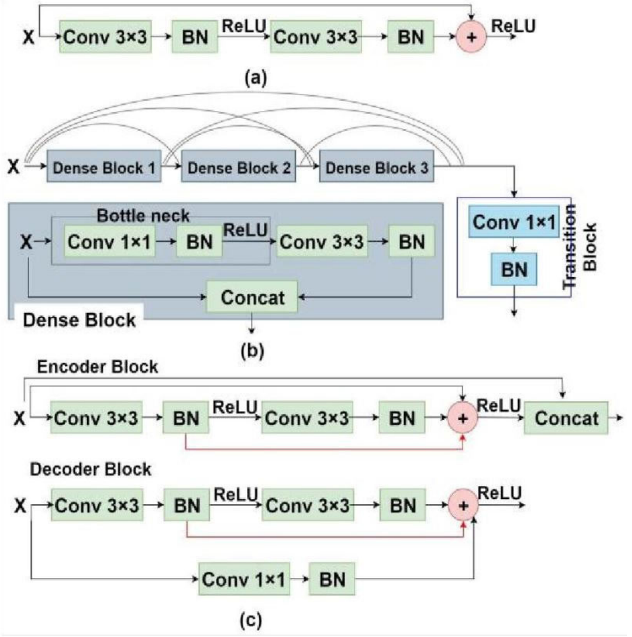

FIGURE 4. **DRUnet architecture implemented for epiphyte segmentation in this study [15]. a) Residual block. b) DenseNet c) The encoder-decoder blocks in DRUnet.**

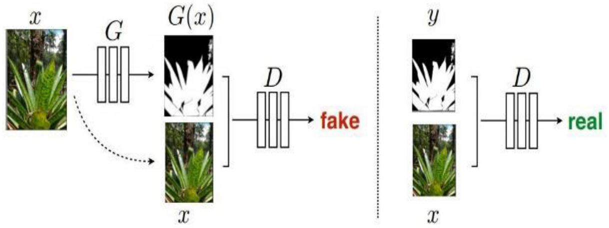

FIGURE 5. **Training a conditional GAN to map epiphytes image to segmented mask. The discriminator, D, learns to classify between fake (produced by the generator) and real {Mask, Epiphyte image} tuples. The generator, G, learns to fool the discriminator.**

The encoder part in the UNet down samples the input image using filters that create feature maps. The Transformer layers convert these feature maps from the encoder into vectorized embedding's using 1 × 1 patches. This study uses a patch size of 16 × 16. Here the vectorized patches are projected into a d-dimensional embedding space using a trainable linear projection. A positional embedding is also added to this to learn the position of the patches. These embedding's are later given to the Transformer layers that are connected in-between for the feature learning. The Transformer encoder

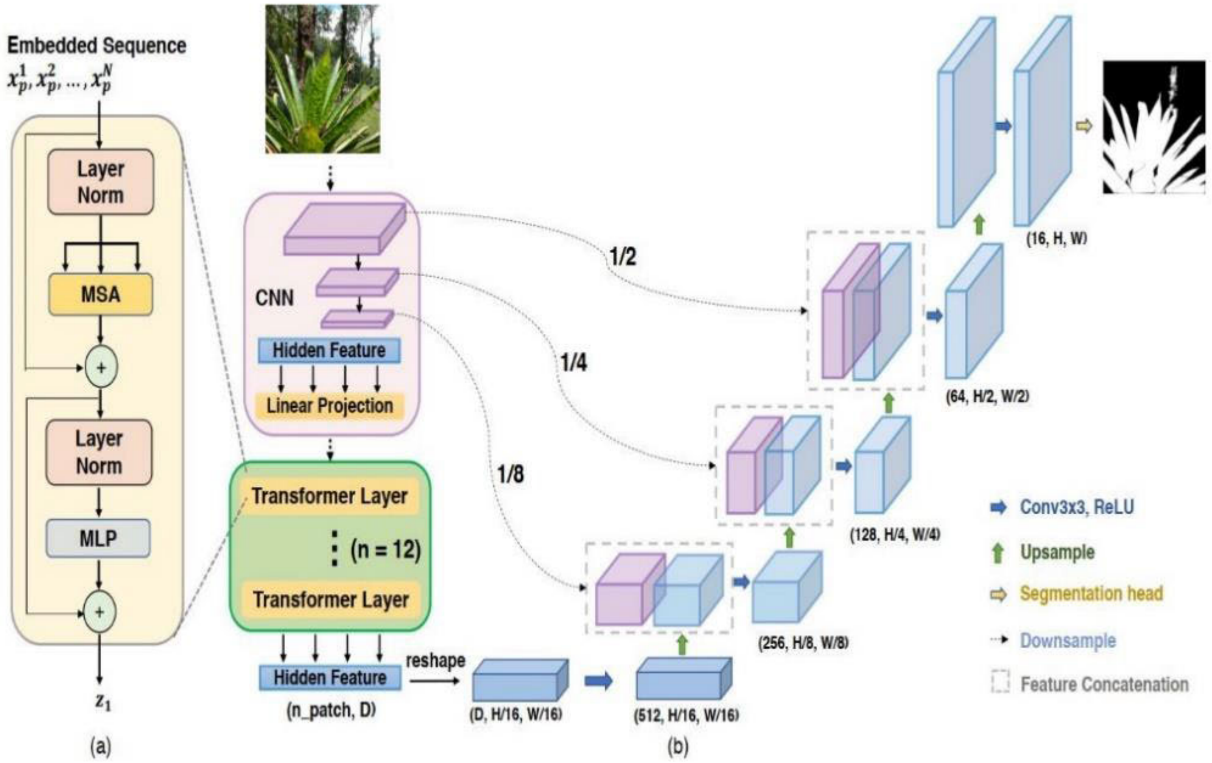

FIGURE 6. Training a) The components of transformer layer. b) The TransUnet layers used for Epiphyte segmentation [25].

consists of Multihead Self-Attention (MSA) and Multi-Layer Perceptron (MLP) blocks. In this study, 12 such layers of Transformer blocks are stacked for learning the hidden features.

The overall architecture of the TransUnet used in this study is depicted in Fig 6. The number of blocks can be considered as a hyper parameter to improve the model performance. The input images converted to a sequence of patches will help the network to learn robust features even though the epiphyte occupancy vary in drone captured images. The study also analyses the model's performance while we have fewer samples for training. The capability of TransUnet to preserve global context and merge high-level features while decoding to preserve local information motivated us to use this model for segmenting a challenging dataset like epiphytes.

IV. RESULTS AND DISCUSSION

The primary challenge with given dataset is fewer training samples and heterogeneous quality. The learning and adaptation of these networks are significant if the model is able to achieve ideal training with high validation accuracy and low validation loss. Apart from learning and adaptation, the model trained with optimum weights is capable to precisely localize the target and preserve the context under varying quality in test images. The presence of Unet in 3 networks and model trained out of only Unet will reveal the significance of additional layers in other 3 models and capability of Unet to deal with fewer training samples. The quantitative performance evaluation of the models generated out of 4 networks for epiphyte segmentation are done using the standard segmentation scores Jaccard / IoU [26]. Apart from the quantitative analysis, the performance of the models while qualitative inferences on the trained models help to analyze how well each model is adapted to deal with challenges in dataset.

A. LEARNING AND ADAPTATION OF SEGMENTATION NETWORKS

The validation loss, accuracy and training loss and accuracy plots obtained by training the four networks for 500 epochs are presented in Fig 7.

The Pix2Pix has a different implementation of loss functions hence those plots appear different than the ones generated from the other three networks. Apart from the loss and accuracy curves, testing with images of varying quality are qualitatively analyzed by visual comparison with ground truth data.

The TransUnet reports only the training accuracy and a dedicated loss function which is combination of binary cross entropy and dice loss. Fig 7a and 7b depict the training, validation accuracy and training, validation loss for Unet model. The Unet model had several fluctuations for the initial epochs while predicting validation data. The disturbances in Unet model settled down after 300 epochs. The spike after 450 epoch shows poor model performance due to few samples in validation set. The Unet model is able to achieve

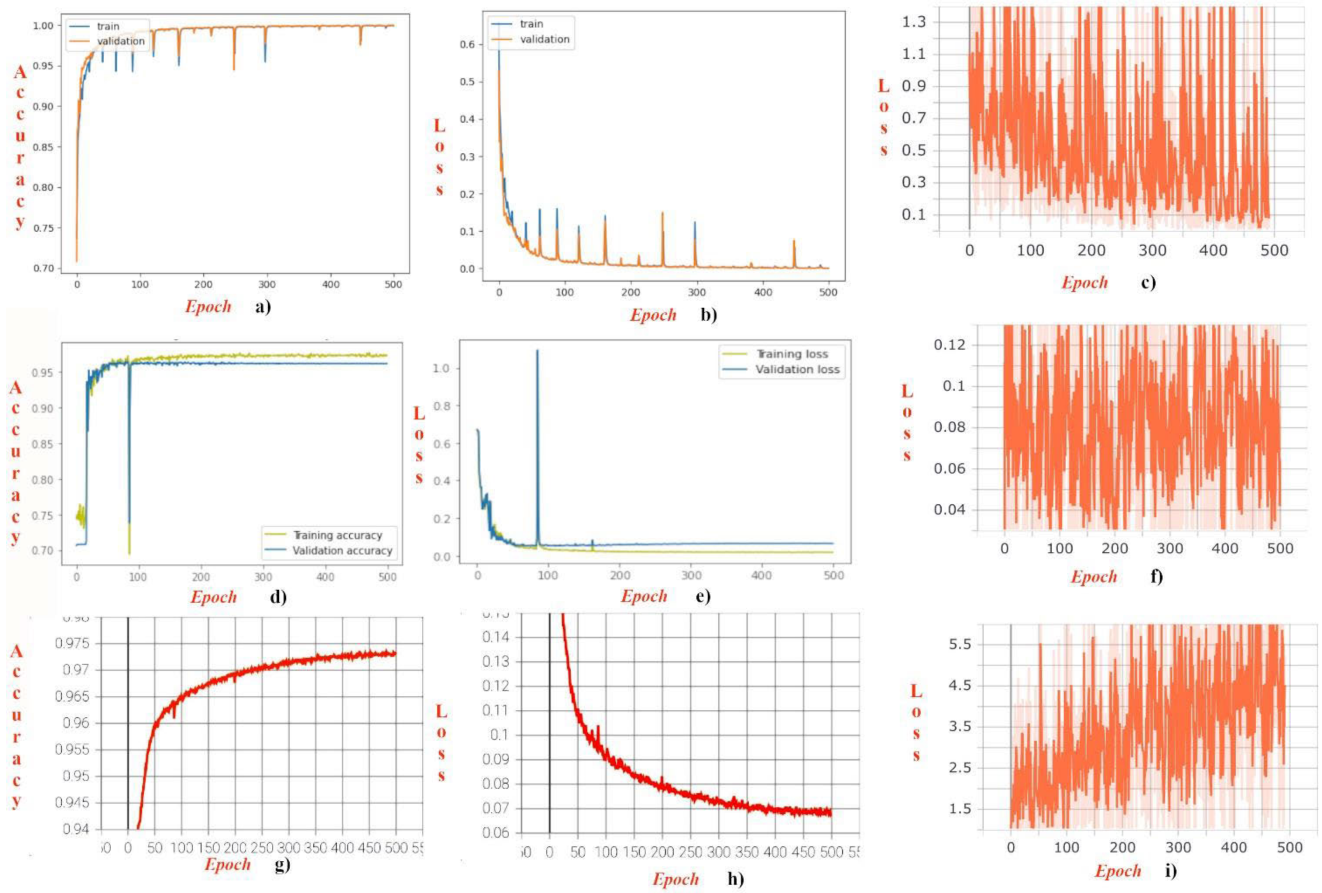

FIGURE 7. Understanding Learning and adaptation of all the 4 networks from validation accuracy and loss plots. (7a, 7b) Validation accuracy and loss for Unet model, (7d, 7e) Validation accuracy and loss for DRUnet model, (7g, 7h) Validation accuracy and loss for TransUnet Model. The pix2pix generator loss (7c) and the discriminator loss (7f) and the total gan loss (7i);.

around 0.97 validation accuracy and 0.05 validation loss after training for 500 epochs.

Figs 7d and 7e plots the validation loss and validation accuracy of DRUnet model for 500 epochs. The validation loss of DRUnet at 100 epoch is very high. Both the validation and training loss settled down for DRUnet after 150th epoch. The DRUnet achieved a minimum validation loss of 0.15 and the highest validation accuracy of 0.95. The TransUnet training accuracy and dedicated loss function values are shown in Figures 7g and 7h, respectively. The performance of TransUnet has typical learning curves with perfect adaptation over iterations. The transitions over the epoch are very smooth, and the growth in scores over the iterations is linear. The transitions in TransUnet loss towards the end of the iterations are settled down, and the changes are not significant after the 450th epoch. TransUnet achieved a highest validation accuracy of. 97 and a minimum validation loss of 0.07 after completing 500 epoch.

Unlike Unet models, DRUnet settled down in 100 epochs and had fewer fluctuations after this. In Fig 7d the training loss trend line is lying below the validation loss and in Fig 7e the training accuracy is above the validation accuracy, this shows the over fitting of the model and leads early decaying of loss.

The DRUnet is a deeper network than Unet and deeper networks over fit while training with less data samples. The frequent spikes in Unet loss and accuracy plots indicate the poor performance of the model for specific samples.

Among the 4 models the transunet model is having an ideal learning curves in Figs 7g and 7h. the smooth transitions and decaying of loss over the epochs show a good model characteristics. the transunet implementation did not had validation loss specifically, rather it had dedicated loss function (binary cross entropy + dice loss) which resulted in training loss and accuracy plots. In an ideal situation, both generator and discriminator will have loss values close to each other. The generator loss shown in Fig 7c is having very high fluctuations and it is varying between 0.1 and 1.3. A similar trend is repeating in discriminator loss shown at Fig 7f and it is varying between 0.04 and 0.12. These frequent changes in Figs 7c and 7f shows the poor performance of generator and discriminator network in this Pix2Pix gan architecture. The plot shown at fig 7i depicts the overall loss function of the Pix2Pix gan architecture. the total loss value for this

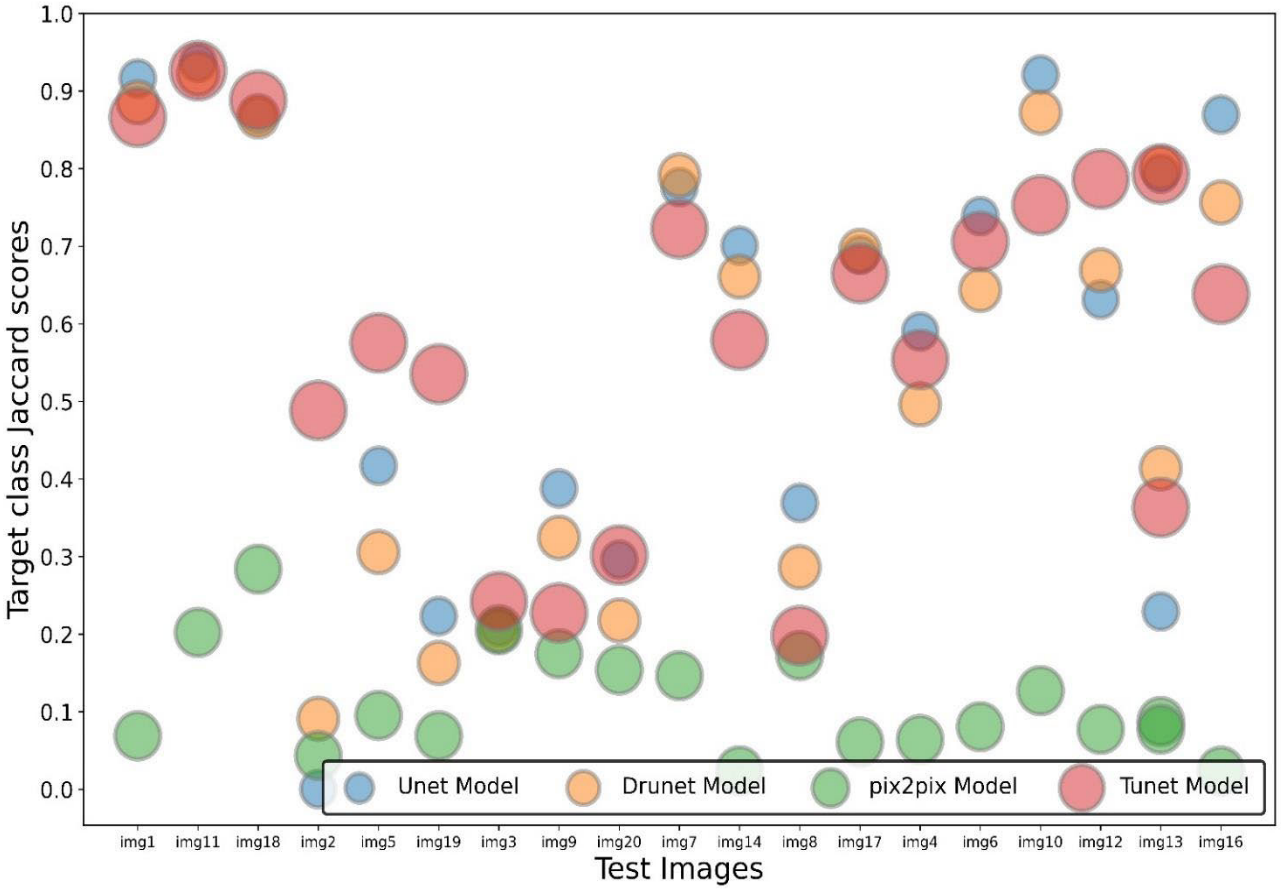

FIGURE 8. The Jaccard segmentation scores by the 4 models towards target class for 20 test images.

Pix2Pix architecture is between 7 and 15. From the learning curves of the models depicted in Figure 7 the transunet is having the well-trained model followed by unet which shows average performance and poor performance. For few selected images the drunet and Pix2Pix shows performance degradation among the 4 models.

B. TARGET CLASS SEGMENTATION PERFORMANCE

This section includes the details of the four networks and their ability to produce accurate labels for 20 test images of heterogeneous quality. Table 2 gives the list of images with each category. The models were trained to generate the labels towards two classes 1) Target class and 2) Background class, as shown in Fig 1. In the predicted images generated by the models, the target and background classes appeared as white and black pixels respectively.

The Jaccard score is computed between the ground truth mask generated by the expert and the predicted mask generated by a trained model. The Jaccard scores range between 0 and 1, a value close to 1 indicates high similarity, and 0 indicates the least similarity between the ground truth and predicted masks.

The bubble plot given in Fig 8 depicts the jaccard scores achieved by the various models while segmenting the epiphyte images. Each trained models can be uniquely identified by the size and color of the bubble. The overlap between the bubbles in the plot indicates a similar jaccard scores achieved by two or more models for the same test image. The bubbles are made transparent such that if multiple models shows a similar performance the overlap between different models will be visible.

TABLE 2. The characteristics of test images among 20 test images.

ID	Image Characteristics	Test Image Number	# of Images
1	Images acquired close to the target epiphyte	1,11 and 18	3
2	Images acquired far away from the target epiphyte	2,5 and 19	3
3	Presence of target epiphyte with visually similar plants	3,9 and 20	3
4	Images acquired under low lighting conditions	7 and 14	2
5	Partially visible target epiphyte	8 and 17	2
6	Images with good lighting conditions and imaged at optimum distance.	4,6,10,12,13, 15 and 16	7

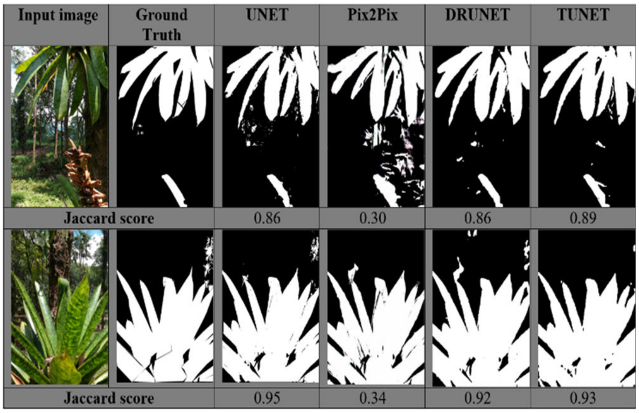

FIGURE 9. The Performance of the models for images captured close to the target epiphyte.

1) TEST PERFORMANCE FOR CATEGORY ID-1

Images in this category were captured close to the target under good lighting conditions. Two sample images from this category and their Jaccard scores for the target class from each model are given in Fig 9. Based on the Jaccard scores for images taken close to the target, all models are well trained to segment the target except Pix2Pix model.

Scores from the Pix2Pix model were affected by the false positives, the edges of the target and the small separation among the leaf of the target plant are not well preserved. Unet can be used when majority of the images contain the target in close proximity and good lighting conditions. TransUnet and DRUnet was also able to learn and well adapt for these types of images.

The overlap between the bubbles in Fig 8 shows a similar performance among the models for certain test images. For the remaining three models except Pix2Pix, performance for image 18 shows high overlap and the TransUnet scores are slightly higher.

A similar performance by 4 models is repeating for image number 11. These images are captured close to the target under good lighting conditions. Since all the models could predict images under Category ID 1 except Pix2Pix model, we recommend Unet model can be used with less training overhead can perform well for these set of images.

2) TEST PERFORMANCE FOR CATEGORY ID-2

Images in this category have the target farther away from the camera and also under poor lighting conditions. This is due to the limitations of flying drones in dense canopy. The TransUnet outperformed all other 3 models for this category of images. The Unet and DRUnet completely failed to predict the target whereas Pix2Pix predictions were affected by false positives. The sample predictions and Jaccard score from each model is given in Fig 10.

When the drones are operating in dense canopy, it is difficult to reach near to the target to be imaged due to obstacles. This will end up in images where the target visibility is poor. The models except Pix2Pix and TransUnet completely failed to make the predictions. The predictions made by the Pix2Pix contains lot of false positives. None of the models except TransUnet can guarantee a proper prediction during this situation. The TransUnet had fewer false positives compared to other models. The predictions made by TransUnet for these images may not preserve the proper shape and structure due to smoothed edges.

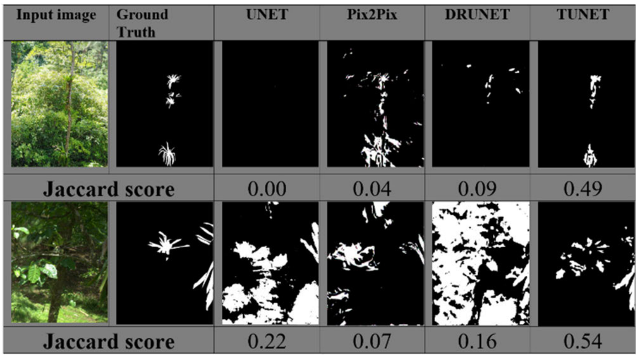

FIGURE 10. Performance of the models for images captured far away from the target.

The hidden CNN features with linear projection to transformers layers as sequence of embedding's in TransUnet helped to acquire minute details of target while imaged from a higher distance. For future missions, we recommend bringing the drone as close to target which will make sure a guaranteed predictions by the models. The predictions done by TransUnet will work in images same as shown above. The TransUnet cannot guarantee proper predictions beyond this distance. The predictions made by TransUnet in higher distance may have smoothed edges with less target details. The presence of Unet encoder-decoder and Transformer layers in TransUnet helped to localise the target precisely and retain the context. This shows that TransUnet is able to learn and adapt for these test cases after training with minimum samples.

3) TEST PERFORMANCE FOR CATEGORY ID-3

Images in this category had other vegetation that appeared similar to the target plant. The performance for these images are shown in Fig 11. The difference in Jaccard scores among the models for these category of images vary less. All 4 models had uniform performance for this category of images. The false positives pixels due to visually similar background is high from all the models. Compared to other 3 models TransUnet had less false positives pixels. This shows that all the four models learned the shape characteristics of the target plant very well which caused the false positives from similar shape plants. The Unet layers in attentional models helped to localise the target plant even though the background vegetation's having visually similar plants. The better performance of TransUnet also shows its capability of context understanding with effective use of embedding's processed at patch size of 16 × 16.

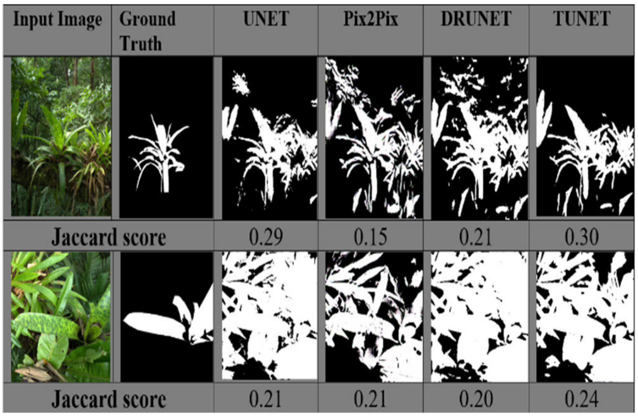

FIGURE 11. The Performance of the models for test images where target occurrence with visually similar plants.

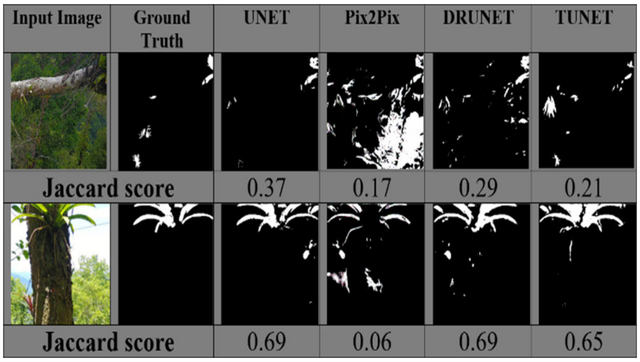

FIGURE 12. Performance of the models for test images where target is partially visible.

The identification of epiphytes in species level will be challenging for all 4 models under this situation. We recommend family level identification rather than species level in test images where target species appears with visually similar species from same family.

4) PERFORMANCE FOR CATEGORY ID-4

The images in this category are poor illuminated. The poor illumination is due to presence of target in dense forest which blocks the sunlight and shades from nearby vegetation's.

Fig 12 shows the performance for these category of images. Unet and DRUnet performed well for target imaged closer and under poor lighting (Fig 12 rows 2 & 3). The images with poor illumination and target farther was predicted well by TransUnet model (Fig 12 row 1).

We recommend Unet / DRUnet when test images are captured close to target with poor illumination. The TransUnet is not able to make good predictions compared to Unet and DRUnet in this scenario. The low contrast images cause the patches generated by the TransUnet to have uniform information across all patches. The TransUnet is processing all patches in 16 × 16 size. Whereas Unet and DRUnet having different size of convolution kernel enables identifying target when the contrast is uniform in images. TRansUnet is highly recommended when you have target captured farther away from the camera under poor / good lighting conditions. Enhancing these images prior to testing may introduce noise which may not appear naturally.

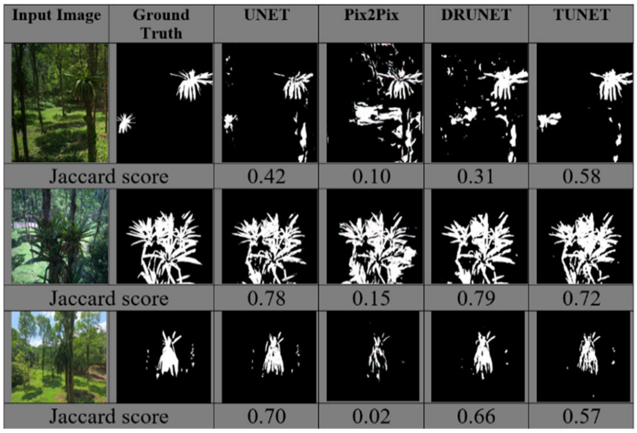

FIGURE 13. The Performance of the models for test images where target under poor lighting / under shade. Row 1- Test image 5 under poor illumination and imaged from higher distance. Row 2 and 3 contains test image 7 and 14 under poor illumination.

5) PERFORMANCE FOR CATEGORY ID-5

The images in this category had the target partially visible in the frame. The Unet and DRUnet equally performed for these images and followed by TransUnet. Compared to other models Pix2Pix model had lowest Jaccard scores. The performance of the model for Category ID 5 where the target is partially visible in the test images are reported in Fig 13.

When the target is partially visible the image contains small portion of the target and some of the relevant information may not be visible. In such images, when the convolution filter found it difficult to discriminate the target class from nearby pixels as they have very small region belongs to target class.

The test images considered under partially visible targets had images where 90% of the target is not visible (Fig 13 –Row1) and target with 50 % visible. While the drones fly in dense canopy this may occur due to other vegetation's. The highest score for this first image in Fig 13 is 0.37 achieved by Unet, but it failed to identify the small target present in the branch.

The DRUnet had more false positives. The TransUnet was not able to predict the target with poor visibility whereas, it was able to identify the small target present in the branch and a visually similar target from same family. This shows the potential of TransUnet to identify targets away from camera if it is fully visible. We recommend in future missions the target occupies as much of the frame as much possible. If the target visibility is more than 50 % in the images except pix2pix all the models will have reasonable performance.

6) PERFORMANCE FOR CATEGORY ID-6

The images in this category were captured under good lighting conditions and at an optimum distance. A few sample predictions from this category are shown in Fig 14. In category

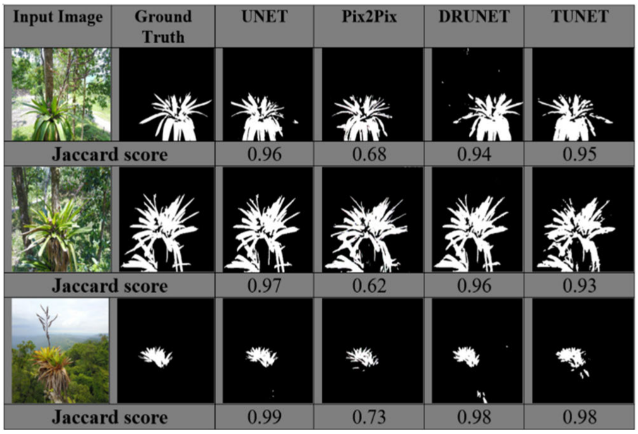

FIGURE 14. Performance of the models for test images captured at optimum distance under good lighting conditions.

ID 1 the target images is fully covering the frame where as in this we will have some portion of the background vegetation also visible. For this category the TransUnet well performed and followed by Unet and DRUnet. The Pix2Pix model is least performed compared to other 3 models. The predictions made by TransUnet is having very good edge information's and also preserved the separation among the leaves compared to Unet and DRUnet.

From the above mentioned inferences after training the 4 segmentation networks, comparing across all 6 categories of test images, TransUnet model showed consistently higher performance in comparison to the other 3 models except for categoires III and V. In Category I and VI Unet outperformed DRUnet and equally performing with TransUnet which could result in saving time during the training step.

The learning and adaptation of the Unet and DRUnet are consistent for categories I, IV and VI. In the TransUnet, the features are processed in a Unet encoder through Transformer blocks as embedding. This resulted in good quality labels from the TransUnet decoder with higher Jaccard scores for the target class. Presence of certain elements of Unet in DRUnet and TransUnet contributed to the localization and context understanding of the target. However, the adaptation and learning of Pix2Pix were not good with fewer images of heterogeneous quality. This caused the Pix2Pix to underperform across all categories of test images. The CGAN based Pix2Pix will not be useful when there are fewer images with heterogeneous quality.

V. CONCLUSION AND RECCOMENDATIONS

Segmentation scores obtained in this study show the potential of Unet and Unet-derived networks to identify target plants when the training samples are relatively less. TransUnet outperformed in comparison to Unet and DRUnet in terms of segmentation scores. The additional layers in TransUnet contributed to better localization and context understanding. However, all four networks struggled to correctly identify the target plant when similar vegetation was present in the images. The optimal performance of Unet across all categories make it ideal for user with minimal hardware resources for training. The TransUnet ability to localize pixels with patches and into embedding's is highly effective for dataset with similar characteristics. However, networks derived from CGAN such as Pix2Pix might not be suitable for identifying target class when the number of images is low and of heterogeneous quality. The individual network layers among 4 models can be considered while building a hybrid model. The post training preprocessing pipelines are possible to enhance the output received from individual models. The individual layers contributions from current networks can be considered for future networks. The effect of augmentation and transfer learning to resolve less data sample can be considered in future study. The real-time implementation of this application in *UAV* embedded platform can consider modified versions of Yolo architectures which can deal with minimum training samples and achieve low latency and high throughput in predictions.

ACKNOWLEDGMENT

The authors would like to thank Professor K.P Soman, Head, Centre for Computational Engineering and Networking (CEN) for all the support and guidance, and CEN students Shashank Anivilla, Anjana K. Menon, and Rahesh Ravi for labeling the epiphytes in the UAV photos used in this study.

REFERENCES

[1] R. Das, V. Pooja, and V. Kanchana, "Detection of diseases on visible part of plant—A review," in *Proc. IEEE Technol. Innov. ICT Agricult. Rural Develop.*, 2017, pp. 42–45, doi: 10.1109/TIAR.2017.8273683.

[2] S. M. Sundara and R. Aarthi, "Segmentation and evaluation of white blood cells using segmentation algorithms," in *Proc. 3rd Int. Conf. Trends Electron. Informat. (ICOEI)*, Apr. 2019, pp. 1143–1146, doi: 10.1109/ICOEI.2019.8862724.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[4] A. Shashank, V. V. Sajithvariyar, V. Sowmya, K. P. Soman, R. Sivanpillai, and G. K. Brown, "Identifying epiphytes in drones photos with a conditional generative adversarial network (C-GAN)," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 2, pp. 99–104, Nov. 2020, doi: 10.5194/isprs-archives-XLIV-M-2-2020-99-2020.

[5] V. V. Sajithvariyar, V. Sowmya, E. A. Gopalakrishnan, P. R. Bupathy, R. Sivanpillai, and G. K. Brown, "Opportunities and challenges of launching UAVs within wooded areas," in *Proc. ASPRS Annu. Conf.*, 2019, pp. 27–31.

[6] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019, doi: 10.1186/s40537-019-0197-0.

[7] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *Proc. Int. Conf. Artif. Neural Netw.*, 2018, pp. 270–279.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[9] J. C. G. Sánchez, M. Magnusson, M. Sandborg, Å. C. Tedgren, and A. Malusek, "Segmentation of bones in medical dual-energy computed tomography volumes using the 3D U-Net," *Phys. Medica*, vol. 69, pp. 241–247, Jan. 2020.

[10] C. Wang, Y. Guo, W. Chen, and Z. Yu, "Fully automatic intervertebral disc segmentation using multimodal 3D U-Net," in *Proc. IEEE 43rd Annu. Comput. Softw. Appl. Conf. (COMPSAC)*, vol. 1, Jul. 2019, pp. 730–739.

[11] J. Owler, B. Irving, G. Ridgeway, M. Wojciechowska, J. McGonigle, and M. Brady, "Comparison of multi-atlas segmentation and U-Net approaches for automated 3D liver delineation in MRI," in *Proc. Annu. Conf. Med. Image Understand. Anal.*, 2019, pp. 478–488.

[12] T. Wang, J. Xiong, X. Xu, M. Jiang, H. Yuan, M. Huang, J. Zhuang, and Y. Shi, "MSU-Net: Multiscale statistical U-Net for real-time 3D cardiac MRI video segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2019, pp. 614–622.

[13] C. Wang, T. MacGillivray, G. Macnaught, G. Yang, and D. Newby, "A two-stage U-Net model for 3D multi-class segmentation on full-resolution cardiac data," in *Proc. Stat. Atlases Comput. Models Heart. Atrial Segmentation LV Quantification Challenges, 9th Int. Workshop, STACOM, Held Conjunct (MICCAI)*. Granada, Spain: Springer, Sep. 2018, pp. 191–199.

[14] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-Net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82031–82057, 2021.

[15] M. Jafari, D. Auer, S. Francis, J. Garibaldi, and X. Chen, "DRU-Net: An efficient deep convolutional neural network for medical image segmentation," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 1144–1148.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020, doi: 10.1145/3422622.

[17] J. Cai and H. Zhu, "Lung image segmentation by generative adversarial networks," in *Proc. Int. Conf. Image Video Process., Artif. Intell.*, Nov. 2019, pp. 175–180.

[18] A. Basu, R. Mondal, S. Bhowmik, and R. Sarkar, "U-Net versus Pix2Pix: A comparative study on degraded document image binarization," *J. Electron. Imag.*, vol. 29, no. 6, Dec. 2020, Art. no. 063019.

[19] H. Tsuda and K. Hotta, "Cell image segmentation by integrating Pix2pixs for each class," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1065–1073.

[20] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[21] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197–207, Apr. 2019.

[22] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[24] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6881–6890.

[25] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[26] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool," *BMC Med. Imag.*, vol. 15, no. 1, pp. 1–28, Dec. 2015.

[27] B. Olimov, K. Sanjar, S. Din, A. Ahmad, A. Paul, and J. Kim, "FU-Net: Fast biomedical image segmentation model based on bottleneck convolution layers," *Multimedia Syst.*, vol. 27, pp. 1–14, Jan. 2021.

[28] B. Olimov, J. Kim, and A. Paul, "REF-Net: Robust, efficient, and fast network for semantic segmentation applications using devices with limited computational resources," *IEEE Access*, vol. 9, pp. 15084–15098, 2021.

[29] B. Olimov, S.-J. Koh, and J. Kim, "AEDCN-Net: Accurate and efficient deep convolutional neural network model for medical image segmentation," *IEEE Access*, vol. 9, pp. 154194–154203, 2021.

[30] P. Sharma, Y. P. S. Berwal, and W. Ghai, "Performance analysis of deep learning CNN models for disease detection in plants using image segmentation," *Inf. Process. Agricult.*, vol. 7, no. 4, pp. 566–574, Dec. 2020.

[31] A. K. Mortensen, M. Dyrmann, H. Karstoft, R. N. Jørgensen, and R. Gislum, "Semantic segmentation of mixed crops using deep convolutional neural network," in *Proc. CIGR-AgEng Conf.*, 2016, pp. 26–29.

[32] S. Sakurai, H. Uchiyama, A. Shimada, D. Arita, and R.-I. Taniguchi, "Two-step transfer learning for semantic plant segmentation," in *Proc. ICPRAM*, 2018, pp. 332–339.

[33] C. Y. N. Norasma, M. A. Fadzilah, N. A. Roslin, Z. W. N. Zanariah, Z. Tarmidi, and F. S. Candra, "Unmanned aerial vehicle applications in agriculture," in *Proc. IOP Conf. Mater. Sci. Eng.*, vol. 506, no. 1. Bristol, U.K.: IOP Publishing, 2019, Art. no. 012063.

V. V. SAJITH VARIYAR received the master's degree in computational engineering and networking (CEN) from Amrita Vishwa Vidyapeetham, India, in 2016, where he is currently pursuing the Ph.D. degree with the Centre for Computational Engineering and Networking. His research interests include AI assisted ecology, agriculture and environment monitoring using unmanned aerial vehicles, and computer vision applications for robotic systems.

V. SOWMYA received the master's degree in remote sensing and wireless sensor networks from the Amrita School of Engineering, Coimbatore, and the Ph.D. degree in artificial intelligence (AI) for natural scene analysis from Amrita Vishwa Vidyapeetham. She joined the Center for Computational Engineering and Networking (CEN), in 2011. Her research interests include AI for signal and image analysis, biomedical, agriculture, and ecology.

RAMESH SIVANPILLAI received the B.Sc. degree in physics from the PSG College of Arts and Science, in 1987, the M.Sc. degree in environmental studies from the Cochin University of Science and Technology, in 1990, the M.Phil. degree in environmental sciences from Bharathiar University, the M.S. degree in environmental sciences from the University of Wisconsin–Green Bay, and the Ph.D. degree in forestry from Texas A&M University, in 2002. He is currently a Senior Research Scientist with the University of Wyoming. His research interests include rapid flood and cropland mapping with satellite images, and object detection in aerial images using AI methods. He was elected as a fellow of the American Society of Photogrammetry and Remote Sensing, in 2021.

GREGORY K. BROWN received the Ph.D. degree from Arizona State University, in 1980. He is currently a Professor with the Department of Botany, University of Wyoming, USA. His specialization is in plant systematics. His research interests include bromeliaceae systematics, with a focus on generic relationships and circumscription in subfamily bromelioideae, morphological and anatomical character mining in bromeliaceae, and bromeliad-based biodiversity.