Fuse-PN: A Novel Architecture for Anomaly Pattern Segmentation in Aerial Agricultural Images

Shubham Innani, Prasad Dutande, Bhakti Baheti, Sanjay Talbar, Ujjwal Baid
Center of Excellence in Signal and Image Processing,
SGGS Institute of Engineering and Technology, Nanded, India-431606
{2016bec035, dutandeprasad, bahetibhakti, sntalbar, baidujjwal}@sggs.ac.in

Abstract

Deep learning and pattern recognition in smart farming has seen rapid growth as a building bridge between crop science and computer vision. One of the important application is anomaly segmentation in agriculture like weed, standing water, cloud shadow, etc. Our research work focuses on aerial farmland image dataset known as Agriculture Vision. We propose to have data fusion of R, G, B, and NIR modalities that enhances the feature extraction and also propose Efficient Fused Pyramid Network (Fuse-PN) for anomaly pattern segmentation. The proposed encoder module is a bottom-up pathway having a compound scaled network and decoder module is a top-down pyramid network enhancing features at different scales having rich semantic features with lateral connections of low-level features. This proposed approach achieved a mean dice similarity score of 0.8271 for six agricultural anomaly patterns of Agriculture Vision dataset and outperforms various approaches in literature.

1. Introduction

Agriculture is an essential sector for income in most of the countries and contributes a significant share to the national economy. For example, In India, the gross domestic product has an 18% contribution from agriculture [29]. According to the Food and Agriculture Organization Unites States (FAO), grain consumption is increasing rapidly and therefore productivity of agriculture must be increased to meet the demands of growing population [40]. This can be mastered by smart farming as a vital element for sustainability and productivity. Smart farming brings application of various technologies in agriculture to increase potential yield and prevent losses [21]. Several challenges in agricultural production include decrease in agricultural lands, climate changes, availability of water resources, presence of weeds and diseases etc [23]. These challenges can be effectively handled by continuous monitoring and analyzing various physical aspects and phenomena.

Remote sensing is widely used for observing and analyzing the agriculture field [32]. The major area involved in aerial agriculture image analysis is segmentation of remote sensing images to generate coarse to fine semantic maps like anomaly segmentation. Anomaly detection refers to the problem of finding various patterns in the data that don't conform to the expected or normal behavior [12]. In this study, several field anomaly patterns are considered that are most important to farmers and have great impact on potential yield of farmlands and it is of utmost importance to accurately locate them. Development of efficient algorithms for detecting field conditions will enable timely actions and planning to prevent major losses and improve yield. However, semantic segmentation of aerial agriculture images is different from other applications due to several aspects like irregular shapes, size and scale, lack of clear boundaries, missng edges, sparsity, absence of texture contrast etc [10]. This makes aerial agriculture image segmentation a unique and challenging task.

While visual pattern recognition on aerial agricultural images carries enormous economic potential, the progress in this area was restricted due to lack of publicly available large scale and high quality dataset for research. The first Agriculture-Vision dataset consisting of RGB+NIR farmland images was proposed in CVPR 2020 for agricultural pattern recognition and to advance the research in this area [10, 11]. We used the publicly available training data in our research and validation data is used as independent test set. Our main contributions are summarized as follows:

- We propose a multi-modal data fusion with RGB and NIR images for anomaly pattern segmentation in agriculture.
- The proposed architecture, Efficient Fused Pyramid

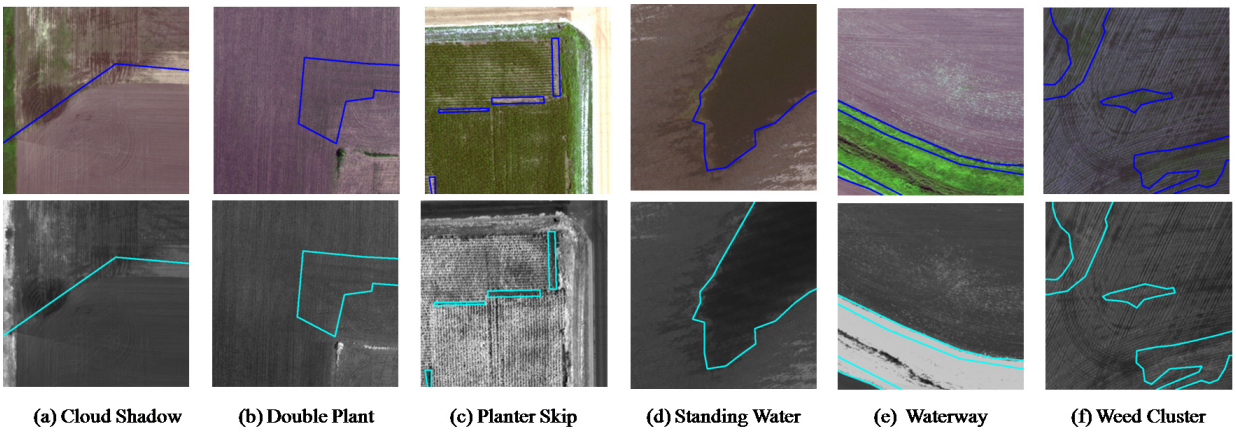

Figure 1: Classwise field patterns from Agriculture Vision dataset are shown from (a) to (f). Various field patterns are marked with dark blue and sky blue in the images.

Network (Fuse-PN) has an encoder module, a bottom-up pathway having a compound scaled network and decoder module, a top-down pyramid network enhancing features at different scales having rich semantic features with lateral connections of low-level features.

2. Literature Review

Many agricultural studies have been proposed in literature to extract meaningful information through images. A review of various machine learning and image processing based methods like Bayesian network, Artificial Neural Network (ANN), Support Vector Machine (SVM), Wiener filter, contrast enhancement etc. applied in agriculture is described in [16, 38]. In recent years, the field of computer vision and pattern recognition has seen tremendous growth with the rise of Convolutional Neural Network (CNN) and deep learning. CNN based approaches have outperformed traditional machine learning based approaches in various tasks like image classification, object detection, segmentation etc. [4, 5, 7]. Deep CNN is used in agriculture vision applications using uav imagery like land cover classification by implying concept of transfer learning [27] and crop classification [35]. A new model was proposed for crop and weed segmentation by ensemble of various CNNs but it was computationally complex [30]. A region growth based algorithm for maize segmentation and R-CNN deep network was proposed in [22]. Few studies worked on discrimination between crops and weeds by using NIR and RGB images [17, 26, 31] and observed that performance is improved with use of such multispectral data.

Aerial image segmentation is a critical problem in the field of visual recognition in agriculture [23]. Typically, algorithms for semantic segmentation of aerial agricultural images are derived from approaches aimed at common semantic segmentation. Fully Convolutional Neural Network (FCNN) [39] and SegNet [1] are among initial attempts of deep learning based algorithms for semantic segmentation and outperformed earlier approaches. The Deeplab series [8] used concept of dilated convolution to enhance the receptive field of neurons and capture multiscale information. SPGNet [9] leveraged multiscale context modules in order to improve performance. A U-shaped fully convolutional network was proposed in [36] especially for medical applications. Feature Pyramid Network [24] was initially proposed for object detection and also proved effective for segmentation. In general, segmentation models mainly consists of two parts: encoder and decoder. Various state-of-art architectures for image classification are typically used for informative feature extraction from the original data. Decoder does the challenging job of upsampling the small resolution feature map to generate precise segmentation map. The research revolves around developing powerful feature extractors (encoders) and upsampling networks (decoders) to improve performance of segmentation [2, 3, 6, 14]

Existing segmentation datasets in literature consist of wide number of object categories that majorly focus on common objects or street scenes [15, 25, 13]. In past few years, some datasets for research in aerial agricultural image visual recognition have been proposed [18, 33]. But they contain less number of high resolution images or less class labels. This gap between agricultural datasets in comparison with common image datasets mentioned earlier hindered growth in this domain [28]. Agriculture Vision is the first publicly available largescale dataset considering high resolution, multi band images and multiple field patterns annotated by agronomy experts. The details of this dataset are described in the next section.

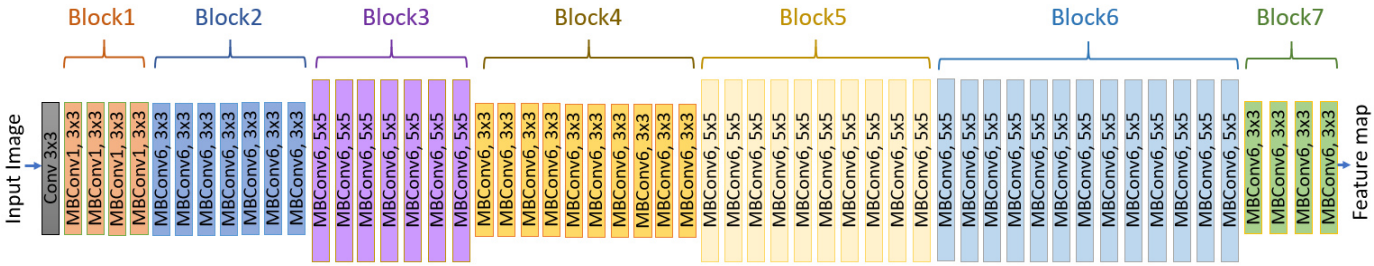

Figure 2: Architecture of EfficientNetB7 as encoder which is build up using MBconv blocks. The overall architecture is divided into seven blocks shown in different colours. Detailed MBConv is shown in Figure 3. Each MBConvX block is shown with the corresponding filter size and the X=1 and X=6 denote the standard ReLU and ReLU6 activation function respectively [3] (©2020 IEEE)

3. Dataset

The Agriculture Vision dataset was proposed in a challenge as part of CVPR 2020 that focused on semantic segmentation of aerial agricultural images [10, 11]. The farmland images were captured between 2017-2019 in growing season and in numerous different farming locations across US. Each image is provided with four input channels viz. Red, Green, Blue (RGB) and Near Infra-Red (NIR). Original captured aerial images were huge in dimension and not suitable for training segmentation models. So, the Agriculture Vision dataset was constructed by cropping field annotations with 512×512 window size.

The dataset consists of six important field pattern annotations viz. cloud shadow, double plant, planter skip, waterway, standing water and weed cluster. These patterns have significant impact on the field conditions and final yield. An example of each class image with ground truth annotation is shown in Figure 1 and ground truth annotations marked as boundaries in each RGB and NIR image. The original dataset was divided into Training (12901 images), Validation (4431 images) and Test (3729 images) set. The dataset also provides boundary maps and six different annotated ground truth as binary masks for training and validation set having six classes. As the challenge is over, validation set is treated as independent test data in performance evaluation.

4. Methodology

In this section, we have briefly illustrated the proposed methodology for semantic segmentation of aerial agricultural images. First the design approach for encoder is described as feature extractor followed by modified multiscale pyramid decoder with lateral connections.

4.1. EfficientNet as Encoder

The semantic segmentation architecures consist of two main modules: Encoder and Decoder. The encoder module

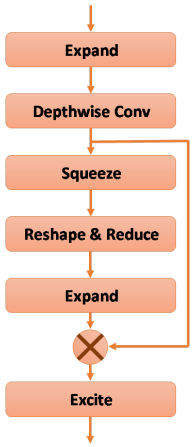

Figure 3: MBConv (mobile inverted bottleneck convolution) basic building block [3](©2020 IEEE)

known as contraction path basically contains a stack of convolutional layers and downsampling blocks to extract dense high-level semantic features from the input image. The encoders are built using various state-of-art CNNs as backbone. We explore effectiveness of EfficientNet which involves idea of scaling depth, width and resolution of network in systematic manner combined with transfer learning. There are eight variants of EfficientNet from EfficientNetB0 to EfficientNetB7 with increasing network depth, width and resolution and also the accuracy. The basic building block of EfficientNet is mobile inverted bottleneck convolution (MBConv) as shown in Figure 3 [37]. It is constructed with depthwise separable convolution, squeeze and excitation optimization [20] and swish activation function [34].

We worked on EfficientNetB7 as encoder whose architecture is shown in Figure 2. This architecture consists of a combination of MBConv modules divided into seven blocks. We modified EfficientNetB7 to incorporate multi-

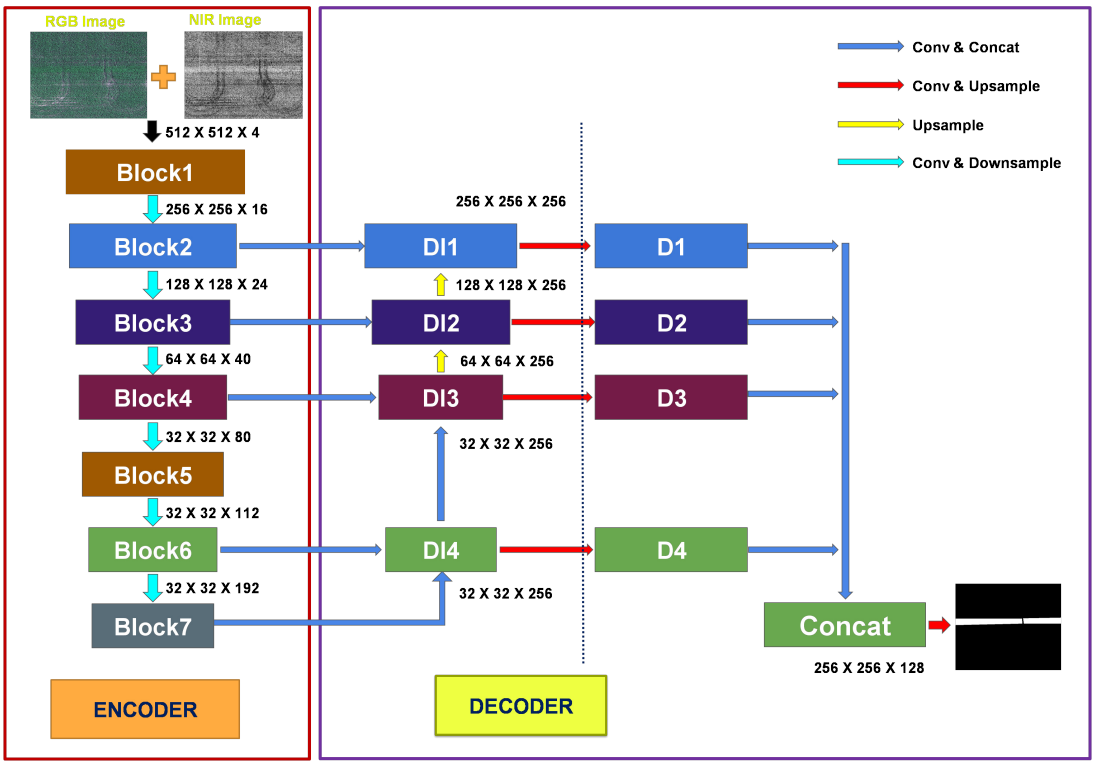

Figure 4: The proposed encoder-decoder architecture for anomaly segmentation. Decoder consists of two stages; DI represents the decoder intermediate block stage while D represents the decoder block stage.

modal input data. Since NIR images are robust to noise with bright textures even in low light conditions, we propose to fuse RGB and NIR modalities and offer them as 4-channel input to the network. Original EfficientNetB7 progressively downsamples the original image such that feature map is 32 times smaller than the input resolution. Such small feature map leads to loss of fine grained details. So, directly using the EfficientNetB7 as encoder doesn't result in adequate performance. To overcome this issue, we propose to use idea of dilated convolution [42] in last few blocks of convolution layers. Thus we maintain the downsampling factor of 16 as well as the receptive field of neurons. This encoder creates a feature hierarchy consisting of feature maps at several scales.

4.2. Feature Pyramid Network as Decoder

The decoder creates higher resolution features by upsampling the spatially coarser and semantically stronger feature maps. In our proposed Fuse-PN architecture, we combine effectiveness dilated EfficientNet as encoder and Feature Pyramid Network (FPN) [24] as decoder. FPN exploits the inherent pyramidal feature hierarchy of deep CNN rather than computing features repeatedly. We build a 2-stage decoder inspired from FPN with modification in Decoder Intermediate (DI) and Decoder (D) blocks as shown in Figure 4. Semantic maps obtained from encoder blocks are first fed to the intermediate decoder block stage. In every block DI, if the spatial resolution of input feature maps is different, feature map of coarse resolution is first upsampled by two using nearest neighbour technique. The feature maps first undergo 1×1 convolution in lateral connection to adjust the number of channels and then the bottom up (low level) and top down (high level) feature maps are merged by element wise addition in the DI block to enhance the intermediate decoder feature maps, as shown in Figure 4. The top down pathway has strong semantic features that is combined with bottom up feature maps having better spatial information via lateral connection.

The output spatial resolution of the finest DI block is half of the size of input image. Rather than directly upsampling it by two, it is passed to Decoder (D) blocks. All the intermediate feature maps from DI1-DI4 block are scaled to the size half of the input size and are then concatenated. It is finally upsurged, followed by 3×3 convolutional layer to obtain the required number of segments in output map. The complete Fuse-PN architecture is tabulated as blocks in Table 1 with operation, output resolution, number of output feature maps, and number of layers. The decoder achieves scale invariance while building high level semantic feature maps as scale change is accounted just by different level in

Table 1: Proposed network – Each row describes a stage i as Block with Operation, Input Resolution, Output Feature Maps and Layers

Block	Operation	Resolution	Channels	Layers
IN	Conv 3x3	512 x 512	1	4
1	MBConv1	256 x 256	16	2
2	MBConv6	128 x 128	24	4
3	MBConv6	64 x 64	40	4
4	MBConv6	32 x 32	80	6
5	MBConv6	32 x 32	112	6
6	MBConv6	32 x 32	192	8
7	MBConv6	32 x 32	256	2
DI	Upsample ->Conv	256 x 256	256	4
D	Conv ->Upsample	256 x 256	128	4
Concatenate	D1+D2+D3+D4 ->Conv	256x256	128	1
OUT	Conv ->Upsample	512 x 512	1	1

the feature pyramid. This is especially useful in this application as the same field pattern can be present in variety of scales in different locations. The Fuse-PN architecture proved itself to be advantageous for segmentation as it allows the use of global context and local information simultaneously.

5. Results and Discussion

The proposed end-to-end semantic segmentation model is built with Tensorflow 2.0 and Keras. The network is trained with the input of 512 $\times$ 512 $\times$ 4 with a batch size of 8 for 100 epochs on NVIDIA P100 GPU. Training was carried out using Adam optimizer with initial learning rate of 0.001. The results are evaluated in terms of the Dice Similarity Score (DSC) which is defined as:

$$Dice = \frac{2 \times TP}{(TP + FP) + (TP + FN)} \quad (1)$$

where TP, FP and FN indicates number of True Positive, False Positive, and False Negative classified pixels.

To accomplish this task of agricultural field pattern segmentation, we choose a multiple CNN approach that segments each class independently. As there exists overlap between the different class labels, Fuse-PN is trained separately for each category. The effectiveness of the data fusion of RGB and NIR modalities with the proposed deep neural network is presented in Figure 5 with result analysis.

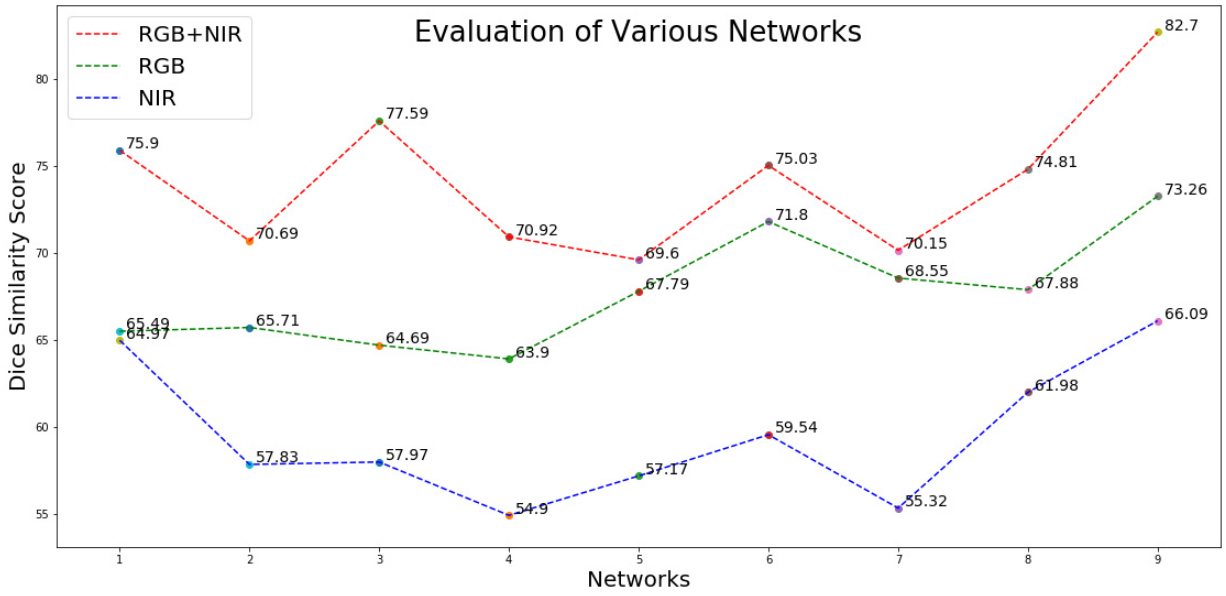

Figure 5: Results on validation dataset from Agriculture Vision using different input modalities with Fuse-PN

Table 2: Classwise Dice Similarity Score of various network architectures with RGB modality

Sr. No.	Network	Classwise Dice Similarity Coefficient						Mean Dice Score
		Cloud Shadow	Double Plant	Planter Skip	Standing Water	Waterway	Weed Cluster	
1	DeeplabV3+ (ResNet50)	64.29	66.68	68.23	72.77	70.08	73.86	69.31
2	DeeplabV3+ (MobileNet)	69.16	62.44	62.6	64.73	66.24	69.06	65.71
3	Res-U-Net	61.59	66.05	67	65.74	66.37	61.36	64.68
4	Mobile-U-Net	65.11	60.95	62.53	61.09	71.25	62.44	63.89
5	Eff-U-Net	63.78	68.85	69.62	73.63	68.68	62.16	67.78
6	Res-FPN	71.42	67.97	75.2	78.1	63.32	74.77	71.79
7	Mobile-FPN	76	62.1	73	66.38	73.02	60.78	68.55
8	Eff-PN	60.98	68.15	63.49	64.66	76.48	73.54	67.88
9	Fuse-PN	66.3	74.15	73.19	76.32	72.96	76.63	73.26

Table 3: Classwise Dice Similarity Score of various network architectures with NIR modality

Sr. No.	Network	Classwise Dice Similarity Coefficient Score						Mean Dice Score
		Cloud Shadow	Double Plant	Planter Skip	Standing Water	Waterway	Weed Cluster	
1	DeeplabV3+ (ResNet50)	67.86	63.34	64.13	62.39	63.97	68.12	64.97
2	DeeplabV3+ (MobileNet)	59.88	52.38	56.48	56.32	57.66	64.25	57.83
3	Res-U-Net	55.92	48.27	58.01	56.89	62.48	66.27	57.97
4	Mobile-U-Net	52.86	55.28	42.85	59.66	54.91	63.82	54.90
5	Eff-U-Net	53.64	56.71	54.97	51.05	60.59	66.08	57.17
6	Res-FPN	58.31	59.96	50.39	57.18	64.31	67.06	59.54
7	Mobile-FPN	57.02	55.31	56.55	60.99	47.2	54.85	55.32
8	Eff-PN	59.68	57.21	59.55	63.33	64.25	67.88	61.98
9	Fuse-PN	63.35	66.23	61.42	66.8	69.1	69.65	66.09

Table 4: Classwise Dice Similarity Score of various network architectures with RGB & NIR modality

Sr. No.	Network	Classwise Dice Similarity Coefficient						Mean Dice Score
		Cloud Shadow	Double Plant	Planter Skip	Standing Water	Waterway	Weed Cluster	
1	DeeplabV3+ (ResNet)	78.43	77.51	78.45	78.51	79.28	81.2	78.90
2	DeeplabV3+ (MobileNet)	67.61	74.17	68.57	67.09	70.05	76.63	70.68
3	Res-U-Net	75.14	78.87	77.02	75.69	77.54	81.28	77.59
4	Mobile-U-Net	71.53	69.53	69.71	71.78	69.69	73.27	69.25
5	Efficient-U-Net	74.81	70.67	71.79	69.57	70.58	75.08	72.08
6	Res-FPN	79.76	69.2	71.14	74.01	77.76	78.29	75.03
7	Mobile-FPN	72.83	70.75	72.81	73.62	68.3	74.57	72.15
8	Efficient-PN	78.64	73.81	79.41	70.32	73.12	73.57	74.81
9	**Fuse-PN**	**85.91**	**81.69**	**79.83**	**82.34**	**81.85**	**84.59**	**82.71**

For comparison of performance, the same model was trained with three different strategies: (i) Only RGB independently, (ii) Only NIR independently and (iii) Fusion of RGB and NIR. As seen from the figure, the dice score of the anomaly patterns like Double Plant, Planter Skip, Weed Cluster, Waterway, Cloud Shadow, and Standing Water, has significantly improved performance with multi-modality compared to the models when trained as an independent modality. The X-axis in Figure 5 represents network from Table 2 while Y-axis represents Dice Similarity Score for a particular modality for the corresponding network.

We compared the proposed Fuse-PN architecture with various other encoder-decoder architectures from literature like DeeplabV3+ [8], UNet [36], FPN [24]. The classwise dice similarity score for various networks evaluated on validation dataset is tabulated in Table 2, 3 and 4 for different combinations of input image modalities. As seen from the tables, the dice score of the anomaly patterns like Double Plant, Planter Skip, Weed Cluster, Waterway, Cloud Shadow, and Standing Water, has significantly improved performance compared to the models when trained as an independent modality. Original UNet consists of a symmetric contraction and expansion path designed for medical image segmentation but its performance was limited. We experimented by changing the contraction path i.e. encoder by different state-of-art CNN networks like ResNet [19], MobileNet [37] and EfficientNet [41] as the backbone in encoder. We also implemented DeepLabV3+ decoder by combining it with ResNet and MobileNet as encoder. After several experimentation with FPN, it was observed that it performed better than other models due to the inclusion of various low level features stacked like a pyramid to form

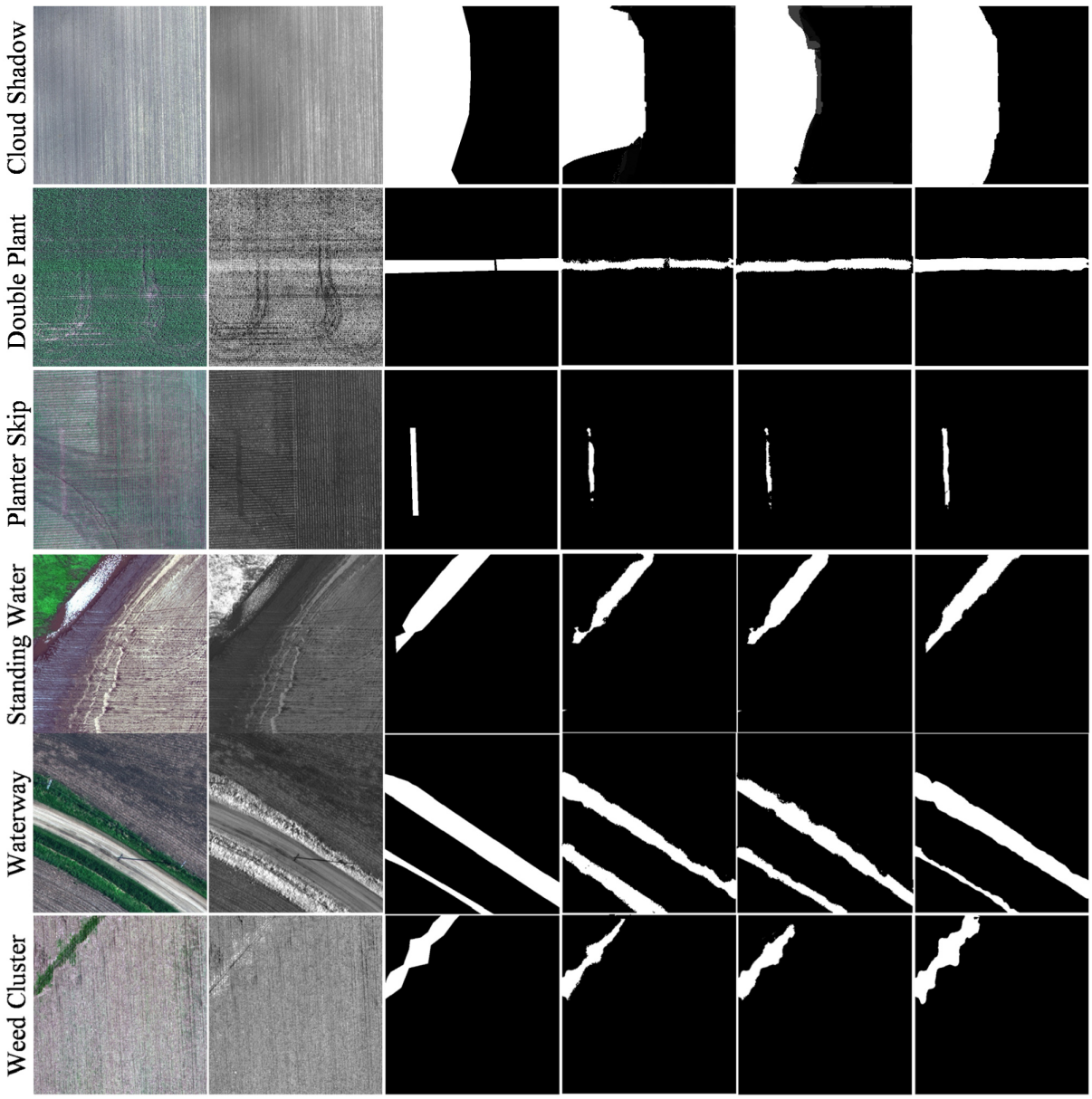

Figure 6: Results on validation dataset from Agriculture Vision. The 1^{st} and 2^{nd} column presents the RGB and NIR image for each pattern respectively. 3^{rd} depicts the labels. The prediction of corresponding samples of NIR modality, RGB modality, Fuse-PN based modality is depicted in column 4^{th}, 5^{th} and 6^{th} column respectively.

the decoder, as shown in Figure 4. These pyramids are useful to extract features and analyze the complex agricultural scenes at multiple levels. Also, alternative lateral connections from the low-level features to the high-level features help the pyramid generate an excellent segmentation map that significantly modifies the original FPN decoder. Results of segmentation for various patterns with training on RGB, NIR, and RGB+NIR modalities are as shown in Figure 6 for visual interpretation. As seen from Figure 6, the model has improved performance when multi-modality is fused. Various approaches are also compared in terms of inference speed on P100 NVIDIA GPU and is tabulated in Table 5. Our Fuse-PN architecture can process 8 images/second at the time of inference.

6. Conclusion

This paper focuses on anomaly segmentation from aerial farmland images using the Agriculture Vision dataset. A novel deep learning architecture is developed to capture the different features of the RGB and NIR images. The

Table 5: Inference Speed for models with multi-modal data fusion.

Network	Inference Speed (Frames per Second)
DeeplabV3+ (ResNet50)	7
DeeplabV3+ (MobileNet)	10
Res-U-Net	9
Mobile-U-Net	13
Efficient-U-Net	16
Res-FPN	8
Mobile-FPN	12
Eff-FPN	14
Fuse-PN	8

proposed deep learning-based Fusion Pyramid Network (Fuse-PN) with multi-modality data of RGB and NIR images performed better than other models. These pyramid features are useful to extract information and analyze the complex agricultural scenes. The mean Dice score with RGB-NIR fusion increased by 20% and 10% respectively in comparison with training on NIR and RGB modality independently. With mean Dice Similarity Score of 0.8271 over 6 classes, Fuse-PN proved to be effective for anomaly segmentation and can be extended to other applications.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:2481–2495, 2016. 2

[2] B. Baheti, S. Gajre, and S. Talbar. Semantic scene understanding in unstructured environment with deep convolutional neural network. In *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, pages 790–795, 2019. 2

[3] Bhakti Baheti, Shubham Innani, Suhas Gajre, and Sanjay Talbar. Eff-unet: A novel architecture for semantic segmentation in unstructured environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2, 3

[4] Bhakti Baheti, Shubham Innani, Suhas Gajre, and Sanjay Talbar. Semantic scene segmentation in unstructured environment with modified deeplabv3+. *Pattern Recognition Letters*, 138:223–229, 2020. 2

[5] B. Baheti, S. Talbar, and S. Gajre. Towards computationally efficient and realtime distracted driver detection with mobilevgg network. *IEEE Transactions on Intelligent Vehicles*, 5(4):565–574, 2020. 2

[6] U. Baid, B. Baheti, P. Dutande, and S. Talbar. Detection of pathological myopia and optic disc segmentation with deep convolutional neural networks. In *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, pages 1345–1350, 2019. 2

[7] Ujjwal Baid, Sanjay Talbar, Swapnil Rane, Sudeep Gupta, Meenakshi H. Thakur, Aliasgar Moiyadi, Nilesh Sable, Mayuresh Akolkar, and Abhishek Mahajan. A novel approach for fully automatic intra-tumor segmentation with 3d u-net architecture for gliomas. *Frontiers in Computational Neuroscience*, 14:10, 2020. 2

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *2018 Europian Conference on Computer Vision (ECCV)*, 2018. 2, 6

[9] Bowen Cheng, Liang-Chieh Chen, Yunchao Wei, Yukun Zhu, Zilong Huang, Jinjun Xiong, Thomas S. Huang, Wen-Mei Hwu, and Honghui Shi. Spgnet: Semantic prediction guidance for scene parsing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[10] M. T. Chiu, X. Xu, K. Wang, J. Hobbs, N. Hovakimyan, T. S. Huang, H. Shi, Y. Wei, Z. Huang, A. Schwing, R. Brunner, I. Dozier, W. Dozier, K. Ghandilyan, D. Wilson, H. Park, J. Kim, S. Kim, Q. Liu, M. C. Kampffmeyer, R. Jenssen, A. B. Salberg, A. Barbosa, R. Trevisan, B. Zhao, S. Yu, S. Yang, Y. Wang, H. Sheng, X. Chen, J. Su, R. Rajagopal, A. Ng, V. T. Huynh, S. Kim, I. Na, U. Baid, S. Innani, P. Dutande, B. Baheti, S. Talbar, and J. Tang. The 1st agriculture-vision challenge: Methods and results. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 212–218, 2020. 1, 3

[11] M. T. Chiu, X. Xu, Y. Wei, Z. Huang, A. G. Schwing, R. Brunner, H. Khachatrian, H. Karapetyan, I. Dozier, G. Rose, D. Wilson, A. Tudor, N. Hovakimyan, T. S. Huang, and H. Shi. Agriculture-vision: A large aerial image database for agricultural pattern analysis. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2825–2835, 2020. 1, 3

[12] Peter Christiansen, Lars N. Nielsen, Kim A. Steen, Rasmus N. Jørgensen, and Henrik Karstoft. Deepanomaly: Combining background subtraction and deep learning for detecting obstacles and anomalies in an agricultural field. *Sensors*, 16(11), 2016. 1

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[14] Prasad Dutande, Ujjwal Baid, and Sanjay Talbar. Lncds: A 2d-3d cascaded cnn approach for lung nodule classification, detection and segmentation. *Biomedical Signal Processing and Control*, 67:102527, 2021. 2

[15] M. Everingham, L. Gool, C. K. Williams, J. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2009. 2

[16] Niketa Gandhi and Leisa J. Armstrong. A review of the application of data mining techniques for decision making in agriculture. *2016 2nd International Conference on Contem-*

porary Computing and Informatics (IC3I), pages 1–6, 2016. 2

[17] S. Haug, A. Michaels, P. Biber, and J. Ostermann. Plant classification system for crop /weed discrimination without segmentation. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1142–1149, March 2014. 2

[18] Sebastian Haug and Jörn Ostermann. A crop/weed field image dataset for the evaluation of computer vision based precision agriculture tasks. In Lourdes Agapito, Michael M. Bronstein, and Carsten Rother, editors, *Computer Vision - ECCV 2014 Workshops*, pages 105–116, Cham, 2015. Springer International Publishing. 2

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 6

[20] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):2011–2023, 2020. 3

[21] N. Jeba, Dr. S C Lingareddy, P. Kowsalya, S. Manju Sree, and S. Swetha. Anomaly detection to enhance crop productivity in smart farming. *International Journal of Pure and Applied Mathematics*, 120:11503–11510, 2018. 1

[22] Shichao Jin, Yanjun Su, Shang Gao, Fangfang Wu, Tianyu Hu, Jin Liu, Wenkai Li, Dingchang Wang, Shaojiang Chen, Yuanxi Jiang, Shuxin Pang, and Qinghua Guo. Deep learning: Individual maize segmentation from terrestrial lidar data using faster r-cnn and regional growth algorithms. *Frontiers in Plant Science*, 9:866, 2018. 2

[23] Andreas Kamilaris and Francesc X. Prenafeta-Boldu. Deep learning in agriculture: A survey. *Comput. Electron. Agric.*, 147:70–90, 2018. 1, 2

[24] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, July 2017. 2, 4, 6

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 2

[26] Philipp Lottes, Markus Hörferlin, Slawomir Sander, and Cyrill Stachniss. Effective vision-based classification for separating sugar beets and weeds for precision farming. *Journal of Field Robotics*, 34(6):1160–1178, 2017. 2

[27] Heng Lu, Xiao Fu, Chao Liu, Long-guo Li, Yuxin He, and Nai-wen Li. Cultivated land information extraction in uav imagery based on deep convolutional neural network and transfer learning. *Journal of Mountain Science*, 14:731–741, 12 2016. 2

[28] Yuzhen Lu and Sierra Young. A survey of public datasets for computer vision tasks in precision agriculture. *Computers and Electronics in Agriculture*, 178:105760, 2020. 2

[29] L Madhusudhan. Agriculture role on indian economy. *Business and Economics Journal*, 6:1–2, 08 2015. 1

[30] C. McCool, T. Perez, and B. Upcroft. Mixtures of lightweight deep convolutional neural networks: Applied to agricultural robotics. *IEEE Robotics and Automation Letters*, 2(3):1344–1351, July 2017. 2

[31] A. Milioto, P. Lottes, and C. Stachniss. Real-Time Blob-Wise Sugar Beets VS Weeds Classification for Monitoring Fields Using Convolutional Neural Networks. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42W3:41–48, Aug. 2017. 2

[32] Florian Mouret, Mohanad Albughdadi, Sylvie Duthoit, Denis Kouamé, Guillaume Rieu, and Jean-Yves Tourneret. Detecting anomalous crop development with multispectral and SAR time series using unsupervised outlier detection at the parcel-level: application to wheat and rapeseed crops. working paper or preprint, Sept. 2020. 1

[33] Alex Olsen, Dmitry Konovalov, Bronson Philippa, Peter Ridd, Jake Wood, Jamie Johns, Wesley Banks, Benjamin Girgenti, O.P. Kenny, James Whinney, Brendan Calvert, Mostafa Rahimi Azghadi, and Ron White. Deepweeds: A multiclass weed species image dataset for deep learning. *Scientific Reports*, 9, 12 2019. 2

[34] Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions. *ArXiv*, abs/1710.05941, 2018. 3

[35] J. Rebetez, H. Satizábal, M. Mota, D. Noll, L. Büchi, M. Wendling, B. Cannelle, A. Pérez-Uribe, and S. Burgos. Augmenting a convolutional neural network with local histograms - a case study in crop classification from high-resolution uav imagery. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2016. 2

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 2, 6

[37] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *CoRR*, abs/1801.04381, 2018. 3, 6

[38] J. P. Shah, H. B. Prajapati, and V. K. Dabhi. A survey on detection and classification of rice plant diseases. In *2016 IEEE International Conference on Current Trends in Advanced Computing (ICCTAC)*, pages 1–8, March 2016. 2

[39] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, April 2017. 2

[40] Arti Singh, Baskar Ganapathysubramanian, Asheesh Kumar Singh, and Soumik Sarkar. Machine learning for high-throughput stress phenotyping in plants. *Trends in Plant Science*, 21(2):110 – 124, 2016. 1

[41] M. Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *ArXiv*, abs/1905.11946, 2019. 6

[42] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *International Conference on Learning Representations (ICLR)*, abs/1511.07122, 2016. 4