Proposing A Deep Learning Based Architecture for Agriculture Vision

Thesis · May 2021

DOI: 10.13140/RG.2.2.26628.86404

3 authors, including:

Anirudh Buvanesh
Birla Institute of Technology and Science Pilani
3 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Soumendu Sinha
Central Electronics Engineering Research Institute
42 PUBLICATIONS **696** CITATIONS
SEE PROFILE

Proposing A Deep Learning Based Architecture for Agriculture Vision

Undergraduate Thesis

Submitted in partial fulfillment of the requirements of
BITS F424T Thesis

By

Anirudh Buvanesh
ID No. 2016B4A70614P

Under the supervision of:

Dr. Pratik Narang
&
Dr. Soumendu Sinha

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI, PILANI CAMPUS
May 2021

Declaration of Authorship

I, Anirudh Buvanesh, declare that this Undergraduate Thesis titled, 'Proposing A Deep Learning Based Architecture for Agriculture Vision' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: $17^{th} May, 2021$

Certificate

This is to certify that the thesis entitled, "*Proposing A Deep Learning Based Architecture for Agriculture Vision*" and submitted by Anirudh Buvanesh ID No. 2016B4A70614P in partial fulfillment of the requirements of BITS F424T Thesis embodies the work done by him under my supervision.

Supervisor
Dr. Pratik Narang
Assistant Professor,
BITS-Pilani, Pilani Campus
Date:

Co-Supervisor
Dr. Soumendu Sinha
Senior Scientist,
CSIR-CEERI, Pilani
Date:

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI, PILANI CAMPUS

Abstract

M.Sc. Mathematics and B.E. Computer Science

Proposing A Deep Learning Based Architecture for Agriculture Vision

by Anirudh Buvanesh

The field of computer vision has seen unprecedented growth since the intoduction of datasets like ImageNet [11], MSCOCO [28]. However, application in the agricultural context has been limited due to both quality and quantity of data that is available. In this work we propose an architecture for segmenting agricultural land images from the Agriculture Vision dataset [8]. We explore the effect of data augmentation using GANs [35] and channel based augmentation using unpaired image to image translation [34] to synthesize additional data channels. We also study the influence of using multiple color spaces for model prediction. Using a multi-color space input and graph convolution based architecture we are able to achieve an MIoU of *0.5630*, which is comparable with existing works and nearly 10% over the baseline model that is based on DeepLabV3+.

Acknowledgements

I would like to offer my special thanks to Dr. Pratik Narang for giving me the opportunity and valuable guidance to pursue my bachelor thesis research. I express my gratitude for his motivation, patience and enthusiasm, that have made this research study an enriching learning experience. I would also like to extend gratitude to my co-supervisor, Dr. Soumendu Sinha for his timely support and consideration. My sincere thanks goes to Computer Science and Information Systems department at BITS Pilani and CEERI Pilani for providing me with this opportunity to get a good exposure to research. Finally, a big thanks goes to my parents for believing in my abilities and supporting my interests.

Contents

List of Figures

List of Tables

Abbreviations

CNN	**C**onvolutional **N**eural **N**etwork
GAN	**G**enerative **A**dversarial **N**etwork
GCN	**G**raph **C**onvolutional **N**etwork
CCNet	**C**riss **C**ross Attention **Net**work
MSCGNet	**M**ultiview **S**elf **C**onstructing **G**raph Convolution **Net**works
HIDeGAN	**H**perspectral-guided **I**mage **D**ehazing **GAN**
GT	**G**round **T**ruth
MIoU	**M**ean **I**ntersection over **U**nion

Chapter 1

Introduction

1.1 An overview of Precision Agriculture

Precision Agriculture is an emerging paradigm for farming which has promised to revolutionize agriculture practices. It advocates the use of monitoring and intervention technologies that help cut down human time and effort, increase crop yield and offer cost savings in the form of resource usage optimization [10]. While precision farming has been around for many years, the last decade has seen it becoming mainstream due to technological advancements and adoption of other, broader technologies. Adoption of mobile devices, access to high speed reliable networks and accurate GPS facilities have been pivotal in characterizing the trend for precision agriculture. Progress in fields like artificial intelligence, computer vision and robotics have fuelled research in this domain. Aerial image semantic segmentation is one of the concerns of agriculture vision due to it's high economic potential. In this study we focus on evaluating existing deep learning based architectures for the aforementioned task and coming up with an architecture for the same.

1.2 Why deep learning?

Since the introduction of ImageNet [11], a large scale image classification dataset, research in computer vision and pattern recognition using deep neural networks has seen great progress [22, 39]. Neural network based algorithms have enjoyed success across multiple domains such as medicine, astronomy, autonomous driving [2, 40, 26], across multiple datasets [12, 15]. The improving performance over the years is mainly attributed to higher amount of data available

across different domains and better compute power which has enabled training of deeper architectures [17, 21]. Deep Learning has catalyzed intelligent management and decision making in many aspects of precision agriculture, such as visual crop categorization [36], real-time plant disease and pest recognition [13], picking and harvesting automatic robots [3].

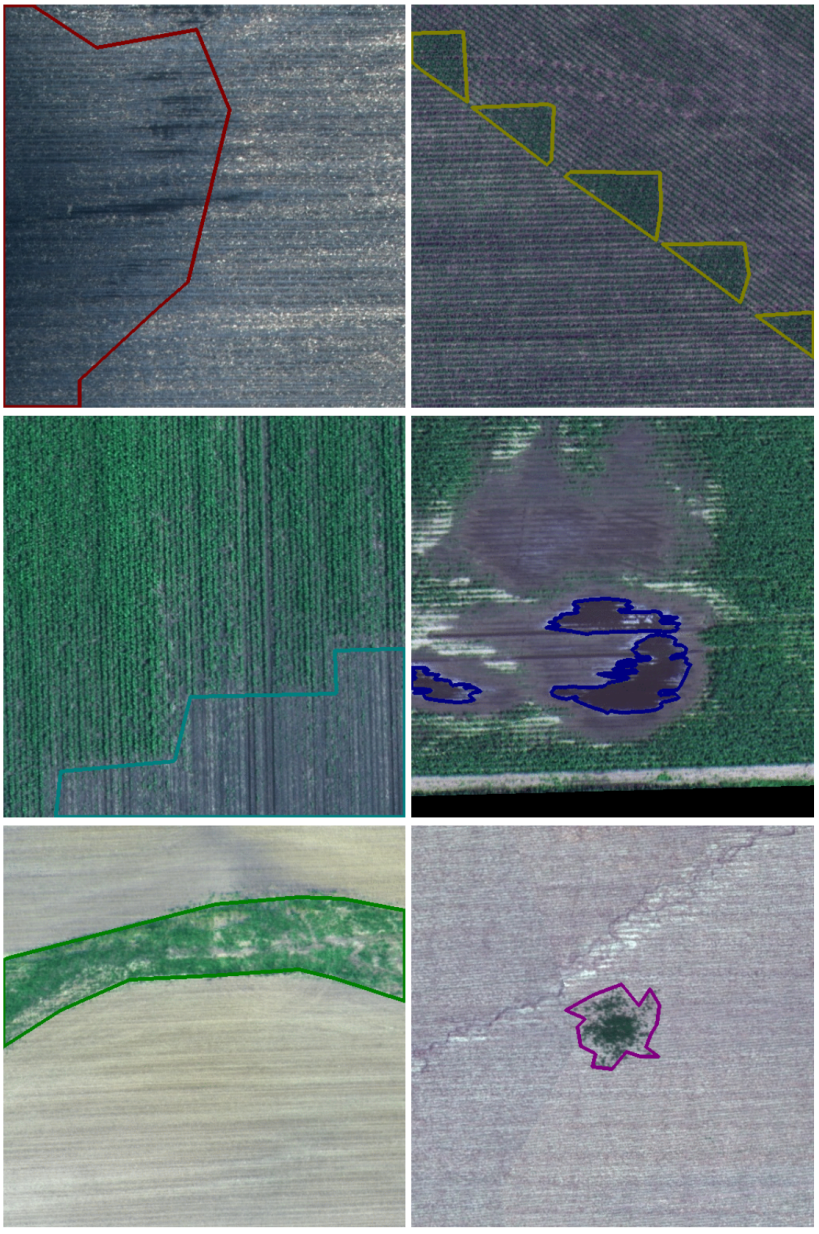


FIGURE 1.1: Segmentation of an agricultural land [8]

1.3 How is vision in Agriculture different?

The essence of deep neural nets achieving state of the art performance in vision tasks such as object detection, classification and segmentation on ImageNet [11] and COCO [28] datasets

lies in the quantum of data available. Scarcity of public image datasets in this domain has been a bottleneck for fast prototyping and evaluation of computer vision and machine learning algorithms. Some of the key challenges here are:

- Procurement of data involves specialized equipment like sensors, UAVs or satellite information (Figure 1.2). Agricultural data is inherently multi-modal, where information such as field temperature and near infra-red signal are important for determining field conditions.
- The task involves detecting and/or segmenting (Figure 1.1) objects such as crops, fruits, weeds which makes the datasets naturally skewed due to low occurrence of classes such as weeds which are of great importance.
- Datasets [5, 8, 38] procured are localized to certain zones which makes generalization of models difficult due to varying appearance of classes in different regions.

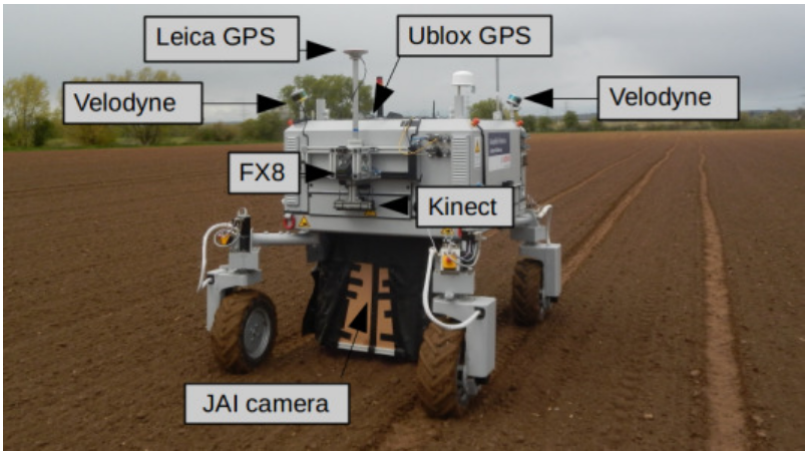


FIGURE 1.2: Setup for Sugar Beet Dataset procurement [5]

Chapter 2

Domain Related Work

In contrast to conventional segmentation tasks that have datasets [46, 9, 28] containing dense annotations for a large number of high resolution images, computer vision on aerial imagery and in particular agricultural context is still in it's infancy. The upcoming sections highlight the datasets we explored and rationale for choosing the *Agriculture Vision Dataset*[8]

2.1 Choice of datasets

Of the datasets explored [31], we narrowed down our focus to a select few (Table 2.1). The rationale for choosing the Agriculture Vision dataset was chosen due to size of dataset, number of places from where data was procured which led to greater variance and the annotations were available for larger number of classes as compared to crop, weed and background.

2.2 Challenges and prior work on Agriculture Vision dataset

Agriculture Vision dataset [8] was launched in 2020, it has 21K images that are divided into 3 parts: train (13K), validation (4K), test (4K). It has annotations for 7 classes: Background, Cloud Shadow, Double plant, Planter skip, Standing water, Waterway, Weed cluster.

Dataset	Dataset size	Channel info.	Features
Agriculture Vision [8]	21K	RGB+NIR	Dataset is unbalanced due to which low accuracy was found in earlier works. Collected from 3K farms across the US. Since the data was collected over a large number of fields it captures a richer variance. Annotations are there for many more fields apart from crop and weed. Primarily consist of Corn and Soybean fields.
Weed Map [38]	10K	Linear combination of RGB+NIR	Data collected from Sugar Beet fields from Switzerland and Germany. Data collection was experimentally collected (i.e seeds sown and crops cultivated over a 5 month period).Has annotation for crop and weed
Sugar Beets [4]	12K	RGB+NIR	Contains annotations for sugar beets and weeds. Images are procured using robots
Joint stem detection [30]	900	RGB+NIR	Has pixel wise annotations for soil, sugar weed, dicot weed, plant weed
CWFID [16]	60	Multi-spectral	The size is very small and the location over which the data has been collected is also very less. Has annotation for crop and weed

TABLE 2.1: Datasets explored for agriculture vision

2.2.1 Multi scale, shape of classes

Classes occur in varying shapes and scales in images (Figure 2.1, 2.2), making it becomes challenging to achieve good results using conventional CNNs which use a single kernel. Architectures have mitigated this issue by making use of spatial pyramid pooling [18], dilated convolutions [6]. The baseline model achieved an *MIoU: 0.434* using DeepLabV3 architecture. Attention based networks [19, 20, 27] have also shown good results on segmentation tasks which have similar problems.

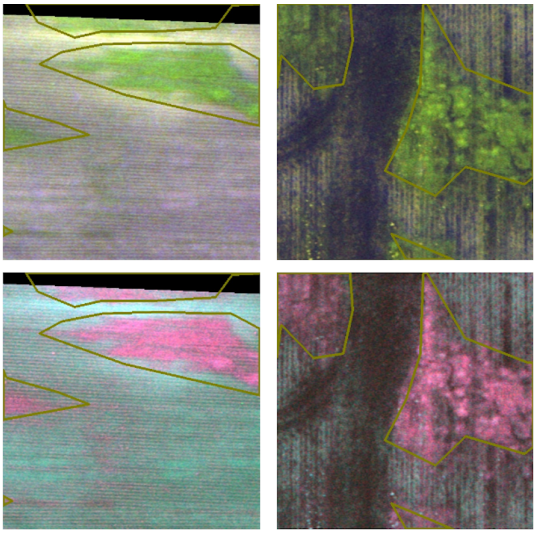


FIGURE 2.1: Weed Cluster

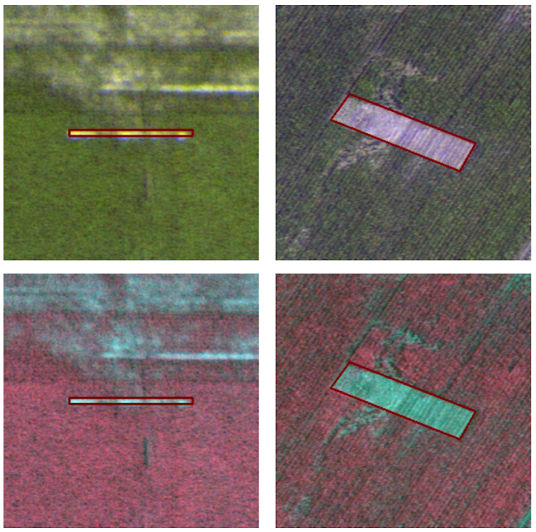


FIGURE 2.2: Planter Skip

2.2.2 Feature divergence

Feature divergence, which refers to difference in appearance between RGB and NIR channels is high (Figure 2.3), for which architectures advocate use of Switchable normalization [42, 32] (SN). SN based architectures have shown an *MIoU: 0.5352*, which is a significant improvement over the baseline model.

2.2.3 Dataset Imbalance

Due to the very nature of the problem there is an inherent imbalance in representation of classes like planter skip, standing water (Table 2.2). To alleviate this issue models make use of augmentation operations like flipping, mirroring and rotation. [29] makes use of dynamic weights in the loss function to give importance to minority classes. Residual DenseNet [7], based on U-Net [37] makes use of an addition network for learning minority class patterns. In this work we explore the effect of using GANs for image synthesis and channel augmentation using different color spaces.

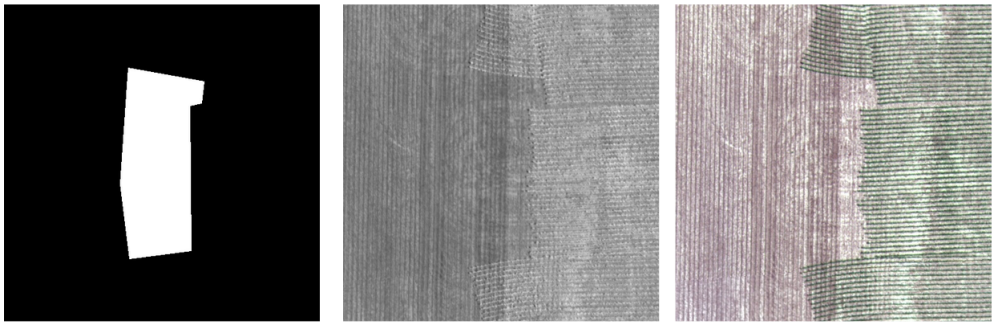


FIGURE 2.3: Feature Divergence between RGB and NIR channels

Class	Number of Images
Cloud Shadow	931
Double Plant	1761
Planter Skip	270
Standing Water	815
Waterway	1769
Weed Cluster	8890

TABLE 2.2: Number of training images containing each annotation class

2.3 Evaluation Metrics

Mean Intersection Over Union (MIoU, Eq. 2.1) is used as the metric for performance evaluation.

$$MIoU = \frac{1}{c} \sum \frac{\text{Area}\,(P_c \cap T_c)}{\text{Area}\,(P_c \cup T_c)} \tag{2.1}$$

where c is the number of classes (here, 7: 6 patterns and 1 background). P_c and T_c are the predicted and ground truth mask respectively.Since it is possible for a single pixel to have multiple annotations a prediction would be correct if the class predicted for that pixel belongs the set of classes with which that pixel has been annotated.

MIoU can also be defined in terms of the confusion matrix, $M^{c \times c}$ as. For each pixel prediction z and it's corresponding annotation set X

$$\text{If } \mathbf{z} \in \mathbf{X},\ \mathrm{M_{x,x}} = \mathrm{M_{x,x}} + 1 \text{ for each } \boldsymbol{x} \text{ in } \mathbf{X}$$
$$\text{Otherwise, } \mathrm{M_{x,x}} = \mathrm{M_{z,x}} + 1 \text{ for each } \mathbf{x} \text{ in } \mathbf{X}$$

Chapter 3

Experiments

3.1 CCNet

Contextual information plays a pivotal role in visual understanding problems. In contrast to FCN which incorporate context through receptive fields of kernels, CCNet [23] proposes an efficient way to incorporate non-local context by stacking a *criss-cross attention module*. State of the art results on Cityscapes test set [9] and ADE20K [46] validation set and GPU memory efficiency were the basis for choosing this model. Due to the class imbalance present in the dataset we explored the alternative of a GAN [14] based augmentation. The augmentation was formulated as an image to image translation problem, with the source domain as the class label map and the target domain as the photo-realistic image, Semantic Image Synthesis with Spatially-Adaptive Normalization (SPADE) [35] was the architecture that was used to learn this mapping. Table 3.2 shows a comparison of ground truth images with generated images that would be used for augmentation. Table 3.1 lists the MIoU obtained on three different schemes: 256×256 image resolution (IR), 512×512 IR, ensemble of the two by taking a maximum over output logits with and without augmentation.

Although results are lagging as compared to baseline *MIoU: 0.434* the performance improvement offered through ensemble model would be vital in discussion of further architectures.

Model scheme	Augmentation	MIoU
256 × 256		0.1720
512 × 512		0.2021
Ensemble		**0.2250**
256 × 256	✓	0.1887
512 × 512	✓	0.1889
Ensemble	✓	**0.2541**

TABLE 3.1: CCNet MIoU on different configurations

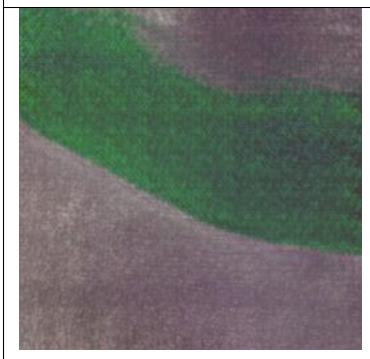
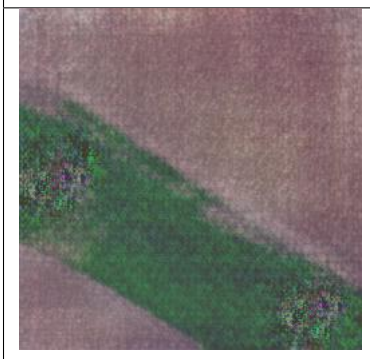
Generated Images	Ground Truth
	
	
	

TABLE 3.2: Comparison of images generated by SPADE [35] with ground truth

3.2 MSCGNet

MSCGNet [29] proposes the use of graph convolutions (GCN) (Appendix A) for segmentation tasks. It uses an attention based variant of ResNeXt [41] as a backbone to learn the adjacency matrix followed by a GCN to predict a segmentation mask. The model achieved an *MIoU: 0.547*. The architecture we propose borrows from MSCGNet and in the upcoming sections we discuss the several modifications to this.

3.2.1 HIDeGAN

HIDeGAN [34] is an architecture proposed for the task of image de-hazing. It takes as input a hazy RGB image and transforms it into a hyper-spectral image (HSI). HSI is then used to as input for another model that maps this image to a dehazed RGB image. Both the transformation maps are learnt through models that are variants of GANs (Appendix A).

Drawing inspiration from HIDeGAN, we propose a channel augmentation of our data (which currently has 4 channels). This is done by learning a mapping from the source domain: Agriculture Vision Dataset [8] to the target domain: MultiPoint dataset [1], it contains 3 channel multi-spectral data in the form of optical thermal (OT) imagery . The mappings $F : NRGB \rightarrow OT$ and $G : OT \rightarrow NRGB$ are learnt by a CycleGAN[47], results of which can be seen in Table 3.3 and Table 3.5. The 3 channels obtained from F were concatenated with NRGB channels and were used for training, this resulted in *MIoU: 0.5073* (Table 3.4)

The degradation in performance as compared to baseline MSCGNet model (*MIoU: 0.547*) was attributed to the unstable training of the GAN whose effect can be seen in Table 3.5, which shows a relatively weak backward mapping from Optical thermal domain to NRGB.

3.2.2 Multiple color space training

Since we weren't able to get performance improvement from learning based approaches, we on training using different color spaces: YCrCb, Lab, HSV. The idea though not novel has shown performance boost in dehazing tasks [33]. Table 3.6 shows the results obtained using SE-ResNeXt-101 and ResNeSt [44] as a backbone with different color spaces [1], the *ensemble* color space refers to taking maximum of the model predictions from all 4 color spaces. For

[1] Architectural tweaks such as addition of couple of layers with residual connections were made to MSCGNet

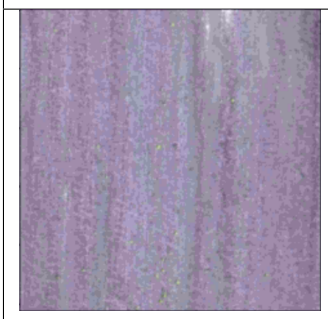
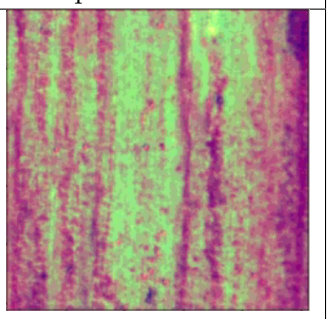
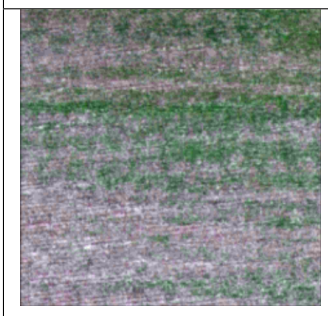
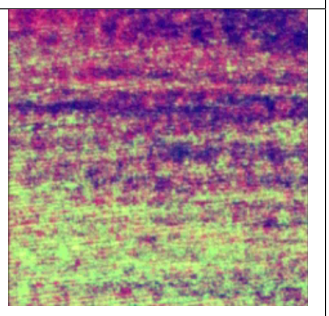
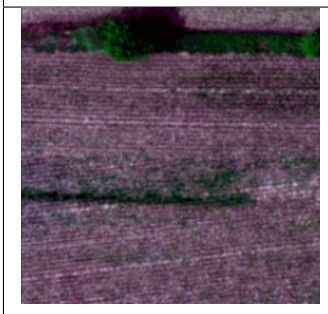
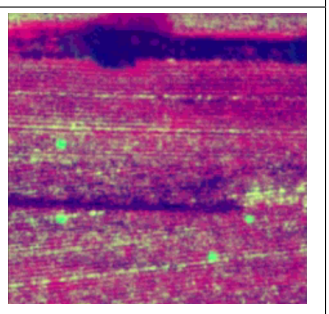
NRGB	Optical Thermal
	
	
	

TABLE 3.3: Mapping NRGB Images to Optical Thermal

Background	Cloud shadow	Double Plant	Planter Skip	Standing Water	Waterway	Weed Cluster	MIoU
0.7994	0.4097	0.5115	0.12289	0.6420	0.5958	0.4705	0.5073

TABLE 3.4: Class-wise IoU for channel based augmentation using Multipoint dataset on validation set

class-wise IoUs refer Appendix B. Figure 3.2 shows a high level representation of the architecture. Table 3.8 shows a comparison between predicted masks and ground truth masks for a few images whose color encoding is given by Figure 3.1. Implementation details can be looked up from Appendix B.

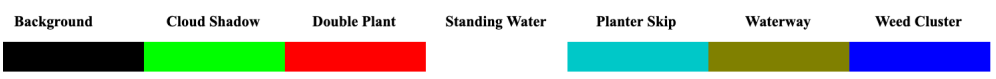


FIGURE 3.1: Mask color palette

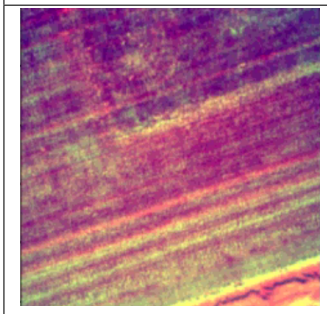
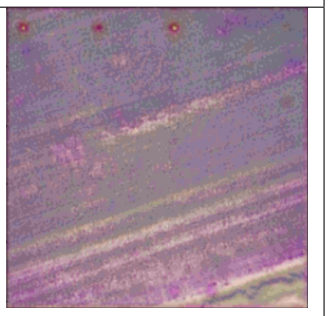
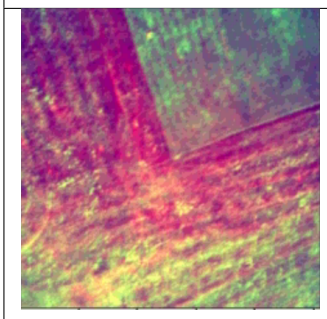
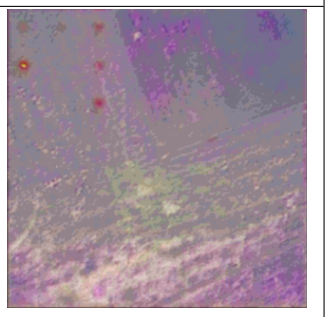
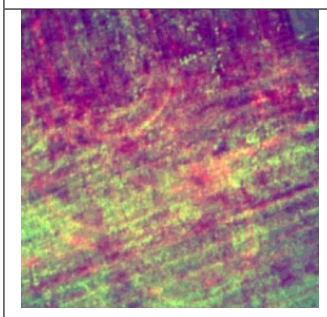
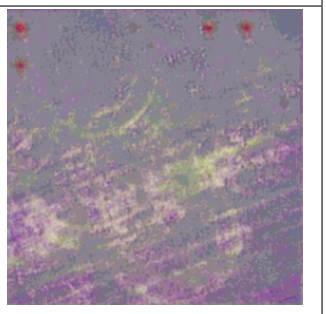
Optical Thermal	NRGB
	
	
	

TABLE 3.5: Mapping Optical Thermal Images to NRGB

Color Space	Backbone	Val. MIoU	Test MIoU
YCrCb + NIR	SE-ResNeXt-101	0.5492	0.5469
RGB + NIR	SE-ResNeXt-101	0.5452	0.5428
HSV + NIR	SE-ResNeXt-101	0.5260	0.5250
Lab + NIR	SE-ResNeXt-101	0.5238	0.5289
Ensemble	SE-ResNeXt-101	**0.5611**	**0.5630**
YCrcb + NIR	ResNeSt-101	0.5430	0.5566
RGB + NIR	ResNeSt-101	0.5509	0.5302
Lab + NIR	ResNeSt-101	0.5329	0.5268
HSV + NIR	ResNeSt-101	0.5350	0.5372
Ensemble	ResNeSt-101	**0.5709**	**0.5473**

TABLE 3.6: Performance with different color spaces using SE-ResNeXt-101 and ResNeSt-101

3.2.3 Fusion block

In the previous section we propose a fusion block which can be used as an extension to the maximum based ensemble approach that was proposed in the previous section, it makes use of channel attention to learn the residual mapping (Figure 3.3) The results obtained till now using the fusion block are comparable to earlier results (Table 3.7)

Background	Cloud shadow	Double Plant	Planter Skip	Standing Water	Waterway	Weed Cluster	MIoU
0.8134	0.5198	0.6130	0.2225	0.6347	0.6683	0.5177	0.5699

TABLE 3.7: Class-wise IoU using fusion block on multi-color space predictions on validation set

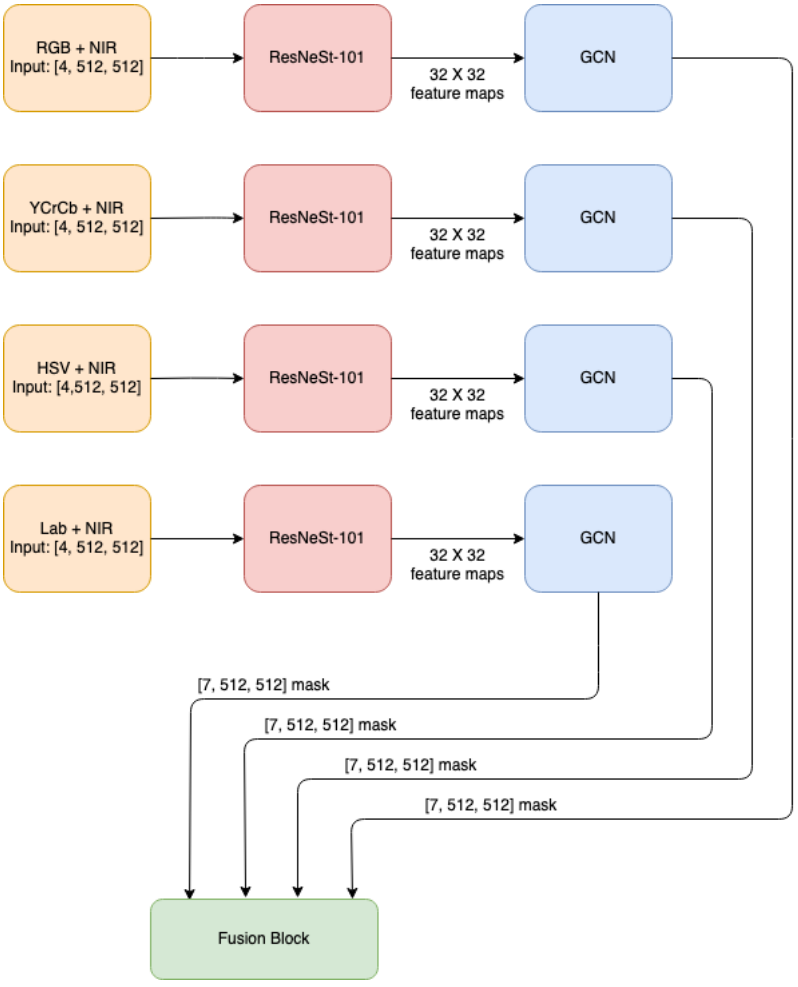


FIGURE 3.2: High level architecture

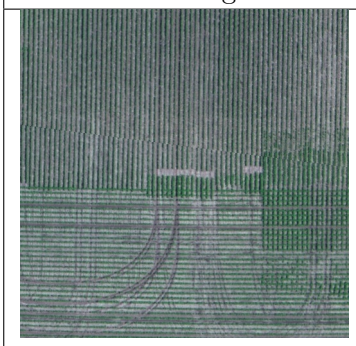
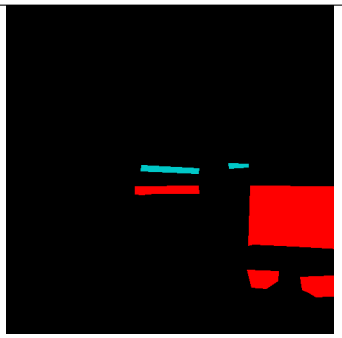
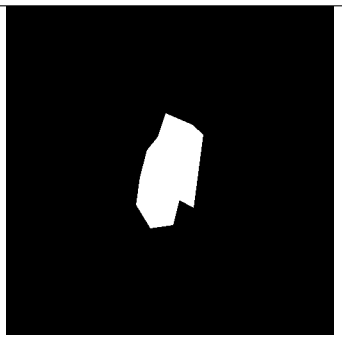
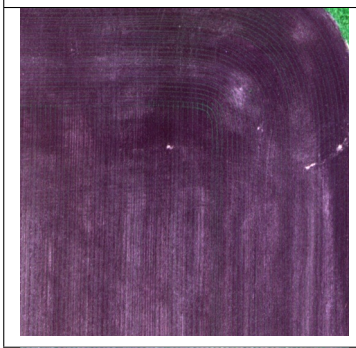
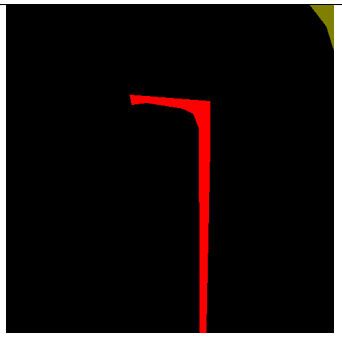
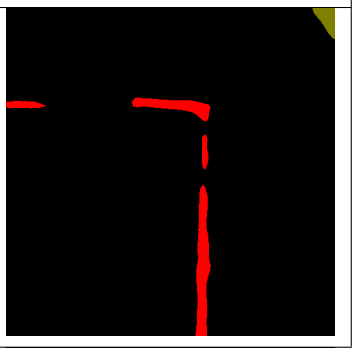
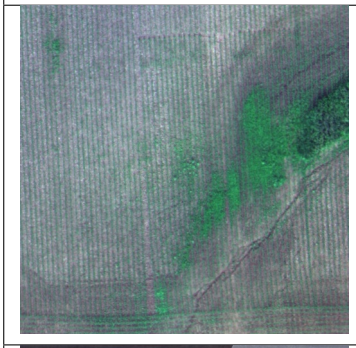
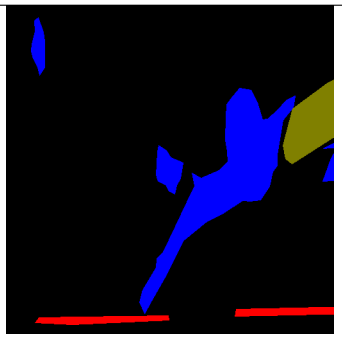
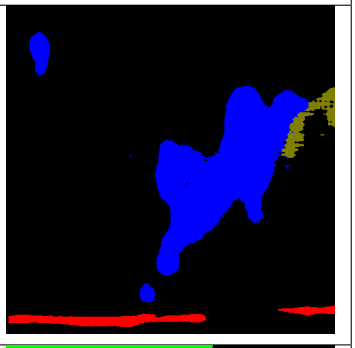
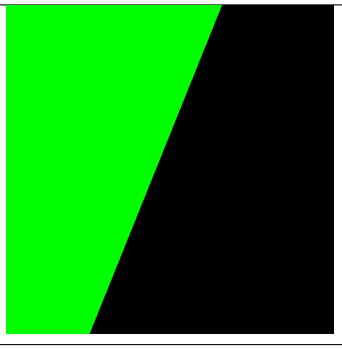
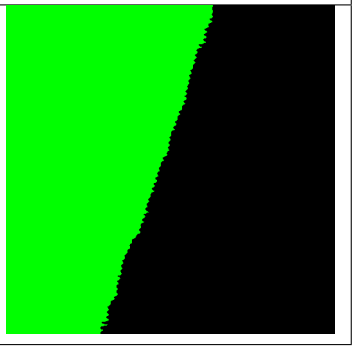
GT Image	GT Mask	Predicted Mask
		
		
		
		
		

TABLE 3.8: Comparison between Ground Truth and Predicted Masks

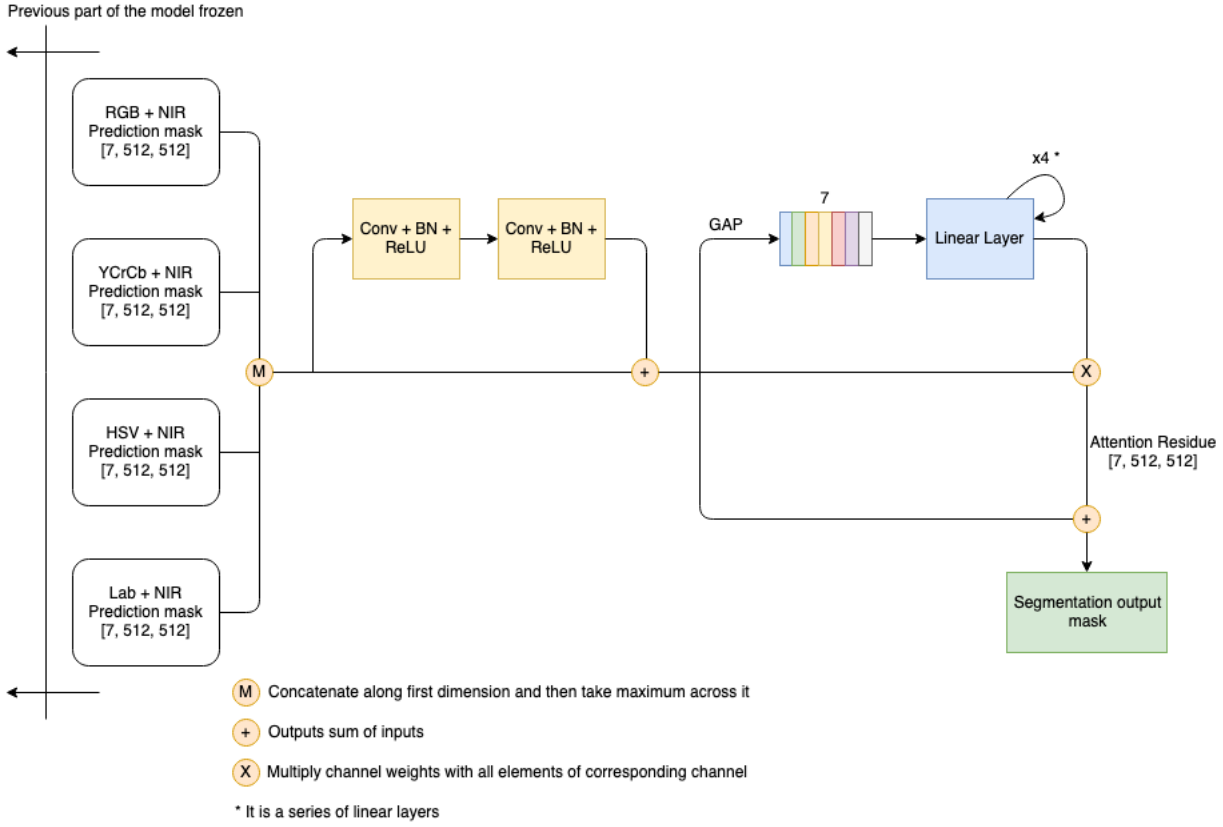


FIGURE 3.3: Fusion Block

3.3 Discussions

The aforementioned architecture achieves a better MIoU of *0.5566* as compared to MSCGNet [29] *0.5470* in the single model case. The study also explores the strategies of combining separately trained models using models trained on different color spaces. The ensemble approach, which combines predictions from different color spaces either through a simple maximum or Fusion Block 3.3 has shown to provide 1-2% MIoU boost (Table 3.6). The best results achieved on the validation and test set are *0.5709* and *0.5630* respectively.

Learning based augmentation strategies, involving data augmentation and channel augmentation were also explored, however results obtained were not as impressive as the former case, which was attributed to unstable training of underlying architecture used for augmentation (Table 3.5 and 3.7)

Appendix A

Common Architectural concepts

A.1 Dilated Convolutions

Dilated convolutions [43] are a type of convolution that inflate the kernel by inserting holes between kernel elements. It helps increase the receptive field of the kernel, keeping the number of parameters same. For example, a dilated 3×3 kernel with a dilation factor of 2 would look at a 5×5 patch in the input. A larger receptive field helps integrate more contextual information, thus making it a good choice for segmentation tasks.

A.2 GAN

GAN [14] is an architecture proposed to learn a data distribution by simultaneously training two models: a generative model G whose task is to synthesize elements which are indifferentiable from the elements sampled from the data distribution and a discriminator mode D which is gets rewarded for correctly identying if the data belongs to the generated distribution or not. The loss function given by A.1 is formulated as a min-max problem and training is done alternatively between generator and discriminator.

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}\,(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))] \tag{A.1}$$

A.3 GCN

GCN [25] are a generalization of convolutional neural networks that are designed to operate on graphs described by an adjacency matrix $A \in \mathbb{R}^{n \times n}$. Each of the n nodes has a feature vector $X_i \in \mathbb{R}^d$. Every node interacts aggregates information from it's neighbours, mathematically the state of the i^{th} node a series of l exchanges is given by $Z_i^{(l+1)}$. Wights are given by $\theta^{(l)} \in \mathbb{R}^{d \times f}$ and initial state $Z^{(0)} = X$. The state update is governed by A.2

$$Z^{(l+1)} = \sigma(AZ^{(l)}\theta^{(l)}) \tag{A.2}$$

A.4 Squeeze Excitation Attention Mechanism

Best described by Figure A.1 is a way of incorporating channel based attention. The block tries to learn weights (importance) that it should be paying to each channel. The resulting output of the block is input scaled by channel weights. Other architectures like Gather Excite Networks [20], CCNet [23] implement feature map attention blocks which weights to quantify the role of other pixels in predicting a specific pixel's output value.

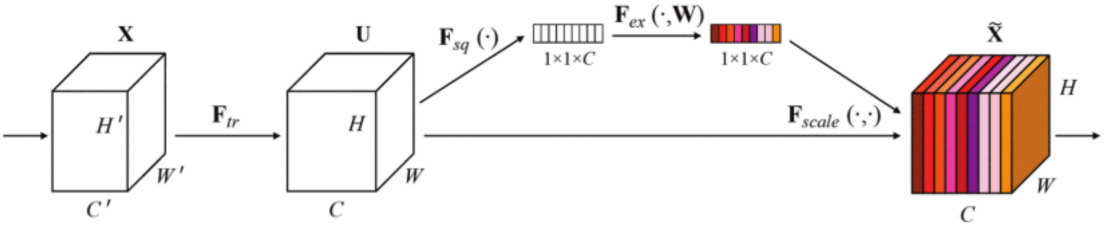


FIGURE A.1: Squeeze Excitation Block [19]

Appendix B

Experimental Details

B.1 Multi color space

Color scheme	Backbone	Backgr.	Cloud shadow	Double Plant	Planter Skip	Standing Water	Waterway	Weed Cluster
YCrCb+NIR	SE-ResNeXt-101	0.8067	0.4902	0.5597	0.1548	0.6507	0.6705	0.5121
RGB+NIR	SE-ResNeXt-101	0.8127	0.4788	0.5540	0.1509	0.6377	0.6693	0.5134
HSV+NIR	SE-ResNeXt-101	0.8016	0.4688	0.5019	0.1439	0.5798	0.6765	0.5097
Lab+NIR	SE-ResNeXt-101	0.8061	0.4611	0.5618	0.0705	0.6011	0.6567	0.5096
Ensemble	ResNeSt-101	0.8121	0.5079	0.5574	0.1688	0.6260	0.6811	0.5199
YCrCb+NIR	ResNeSt-101	0.8026	0.4590	0.5834	0.1579	0.6223	0.6723	0.5038
RGB+NIR	ResNeSt-101	0.8051	0.4723	0.5959	0.2325	0.6111	0.6413	0.4981
HSV+NIR	ResNeSt-101	0.8047	0.4703	0.5712	0.1032	0.6123	0.6536	0.5152
Lab+NIR	ResNeSt-101	0.8087	0.4793	0.5696	0.1680	0.5600	0.6715	0.4929
Ensemble	ResNeSt-101	0.8119	0.5185	0.6119	0.2230	0.6392	0.6692	0.5226

TABLE B.1: Class-wise IoU for different color spaces on validation set

Color scheme	Backbone	Backgr.	Cloud shadow	Double Plant	Planter Skip	Standing Water	Waterway	Weed Cluster
YCrCb+NIR	SE-ResNeXt-101	0.7917	0.4578	0.5309	0.2030	0.6655	0.6418	0.5379
RGB+NIR	SE-ResNeXt-101	0.8023	0.4732	0.5498	0.1550	0.6676	0.6176	0.5345
HSV+NIR	SE-ResNeXt-101	0.7915	0.4354	0.4425	0.1923	0.6673	0.6050	0.5412
Lab+NIR	SE-ResNeXt-101	0.7912	0.4531	0.4669	0.2092	0.6481	0.6076	0.5264
Ensemble	ResNeSt-101	0.8045	0.5025	0.5534	0.2330	0.6578	0.6345	0.5549
YCrCb+NIR	ResNeSt-101	0.7971	0.4137	0.5430	0.2596	0.6659	0.6812	0.5536
RGB+NIR	ResNeSt-101	0.7928	0.3920	0.5590	0.1548	0.7029	0.58114	0.5283
HSV+NIR	ResNeSt-101	0.7935	0.4115	0.5082	0.2452	0.6567	0.6160	0.5295
Lab+NIR	ResNeSt-101	0.7989	0.4296	0.5578	0.0914	0.6396	0.6274	0.5431
Ensemble	ResNeSt-101	0.8005	0.4452	0.5749	0.1552	0.6744	0.6333	0.5477

TABLE B.2: Class-wise IoU for different color spaces on test set

B.2 Implementation details

Backbone architectures (SE-ResNeXt-101 and ResNeSt-101) used were pretrained on ImageNet dataset. The ensemble model makes use of 4 separately trained models using different color spaces as input. Training was done for 19 epochs and 50 epochs on the aforementioned backbones with a batch-size of 10. We made use of Adam [24] and Lookahead [45] as optimizer. The loss function: Adaptive class weighting loss [29] helps mitigate the issue of class imbalance by making use of dynamic weights instead of using pre-computed which are computed via pixel representation of each class in the training data. In case of the fusion block the the entire network (Figure 3.2) except the fusion block was frozen All the training was done on a single NVIDIA Tesla V100-SMX2 GPU.

Bibliography

[1] Florian Achermann et al. "MultiPoint: Cross-spectral registration of thermal and optical aerial imagery". In: *CoRL 2020*. Nov. 2020. URL: https://www.microsoft.com/en-us/research/publication/multipoint-cross-spectral-registration-of-thermal-and-optical-aerial-imagery/.

[2] A. K. Aniyan and K. Thorat. "Classifying Radio Galaxies with the Convolutional Neural Network". In: *The Astrophysical Journal Supplement Series* 230.2 (June 2017), p. 20. ISSN: 1538-4365. DOI: 10.3847/1538-4365/aa7333. URL: http://dx.doi.org/10.3847/1538-4365/aa7333.

[3] Shivaji Bachche. "Deliberation on Design Strategies of Automatic Harvesting Systems: A Survey". In: *Robotics* 4.2 (2015), pp. 194–222. ISSN: 2218-6581. DOI: 10.3390/robotics4020194. URL: https://www.mdpi.com/2218-6581/4/2/194.

[4] Nived Chebrolu et al. "Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields". In: *The International Journal of Robotics Research* (2017). DOI: 10.1177/0278364917720510.

[5] Nived Chebrolu et al. "Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields". In: *The International Journal of Robotics Research* 36.10 (2017), pp. 1045–1052. DOI: 10.1177/0278364917720510. eprint: https://doi.org/10.1177/0278364917720510. URL: https://doi.org/10.1177/0278364917720510.

[6] Liang-Chieh Chen et al. *Rethinking Atrous Convolution for Semantic Image Segmentation*. 2017. arXiv: 1706.05587 [cs.CV].

[7] Mang Chiu et al. *The 1st Agriculture-Vision Challenge: Methods and Results*. Apr. 2020.

[8] Mang Tik Chiu et al. "Agriculture-Vision: A Large Aerial Image Database for Agricultural Pattern Analysis". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020.

[9] Marius Cordts et al. "The Cityscapes Dataset for Semantic Urban Scene Understanding". In: *CoRR* abs/1604.01685 (2016). arXiv: 1604.01685. URL: http://arxiv.org/abs/1604.01685.

[10] M. De Clercq, A. Vats, and A. Biel. "Agriculture 4.0: the future of farming technology". In: *Proceedings of the World Government Summit, Dubai, UAE*. 2018.

[11] Jia Deng et al. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*. Vol. 00. June 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848. URL: https://ieeexplore.ieee.org/abstract/document/5206848/.

[12] Yang Fu et al. "Self-Similarity Grouping: A Simple Unsupervised Cross Domain Adaptation Approach for Person Re-Identification". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019.

[13] Alvaro Fuentes et al. "A Robust Deep-Learning-Based Detector for Real-Time Tomato Plant Diseases and Pests Recognition". In: *Sensors* 17.9 (2017). ISSN: 1424-8220. DOI: 10.3390/s17092022. URL: https://www.mdpi.com/1424-8220/17/9/2022.

[14] Ian J. Goodfellow et al. "Generative Adversarial Nets". In: NIPS'14. Montreal, Canada: MIT Press, 2014.

[15] Yunhui Guo et al. "SpotTune: Transfer Learning Through Adaptive Fine-Tuning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

[16] Sebastian Haug and Jörn Ostermann. "A Crop/Weed Field Image Dataset for the Evaluation of Computer Vision Based Precision Agriculture Tasks". In: *Computer Vision - ECCV 2014 Workshops*. Ed. by Lourdes Agapito, Michael M. Bronstein, and Carsten Rother. Cham: Springer International Publishing, 2015, pp. 105–116. ISBN: 978-3-319-16220-1.

[17] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385.

[18] Kaiming He et al. "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition". In: *Lecture Notes in Computer Science* (2014), pp. 346–361. ISSN: 1611-3349. DOI: 10.1007/978-3-319-10578-9_23. URL: http://dx.doi.org/10.1007/978-3-319-10578-9_23.

[19] Jie Hu, Li Shen, and Gang Sun. "Squeeze-and-Excitation Networks". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7132–7141. DOI: 10.1109/CVPR.2018.00745.

[20] Jie Hu et al. *Gather-Excite: Exploiting Feature Context in Convolutional Neural Networks*. 2019. arXiv: 1810.12348 [cs.CV].

[21] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. "Densely Connected Convolutional Networks". In: *CoRR* abs/1608.06993 (2016). arXiv: 1608.06993. URL: http://arxiv.org/abs/1608.06993.

[22] Gao Huang et al. "Convolutional Networks with Dense Connectivity". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019), pp. 1–1. DOI: 10.1109/TPAMI.2019.2918284.

[23] Zilong Huang et al. *CCNet: Criss-Cross Attention for Semantic Segmentation*. 2020. arXiv: 1811.11721 [cs.CV].

[24] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG].

[25] Thomas N. Kipf and Max Welling. "Semi-Supervised Classification with Graph Convolutional Networks". In: *CoRR* abs/1609.02907 (2016). arXiv: 1609.02907. URL: http://arxiv.org/abs/1609.02907.

[26] David B. Larson et al. "Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs". In: *Radiology* 287.1 (2018). PMID: 29095675, pp. 313–322. DOI: 10.1148/radiol.2017170236. eprint: https://doi.org/10.1148/radiol.2017170236. URL: https://doi.org/10.1148/radiol.2017170236.

[27] Xiang Li et al. *Selective Kernel Networks*. 2019. arXiv: 1903.06586 [cs.CV].

[28] Tsung-Yi Lin et al. "Microsoft COCO: Common Objects in Context". In: *Computer Vision – ECCV 2014*. Ed. by David Fleet et al. Cham: Springer International Publishing, 2014, pp. 740–755. ISBN: 978-3-319-10602-1.

[29] Qinghui Liu et al. "Multi-View Self-Constructing Graph Convolutional Networks With Adaptive Class Weighting Loss for Semantic Segmentation". In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2020.

[30] Philipp Lottes et al. "Joint Stem Detection and Crop-Weed Classification for Plant-specific Treatment in Precision Farming". In: *CoRR* abs/1806.03413 (2018). arXiv: 1806.03413. URL: http://arxiv.org/abs/1806.03413.

[31] Yuzhen Lu and Sierra Young. "A survey of public datasets for computer vision tasks in precision agriculture". In: *Computers and Electronics in Agriculture* 178 (2020), p. 105760. ISSN: 0168-1699. DOI: https://doi.org/10.1016/j.compag.2020.105760. URL: https://www.sciencedirect.com/science/article/pii/S0168169920312709.

[32] Ping Luo et al. *Switchable Normalization for Learning-to-Normalize Deep Representation.* 2019. arXiv: 1907.10473 [cs.CV].

[33] Aryan Mehra et al. "ReViewNet: A Fast and Resource Optimized Network for Enabling Safe Autonomous Driving in Hazy Weather Conditions". In: *IEEE Transactions on Intelligent Transportation Systems* (2020), pp. 1–11. DOI: 10.1109/TITS.2020.3013099.

[34] Aditya Mehta et al. "HIDeGan: A Hyperspectral-guided Image Dehazing GAN". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).* 2020, pp. 846–856. DOI: 10.1109/CVPRW50498.2020.00114.

[35] Taesung Park et al. "Semantic Image Synthesis with Spatially-Adaptive Normalization". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2019.

[36] D. I. Patrício and R. Rieder. "Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review". In: *Comput. Electron. Agric.* 153 (2018), pp. 69–81.

[37] 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* (May 2015).

[38] Inkyu Sa et al. "WeedMap: A Large-Scale Semantic Weed Mapping Framework Using Aerial Multispectral Imaging and Deep Neural Network for Precision Farming". In: *Remote Sensing* 10.9 (2018). ISSN: 2072-4292. DOI: 10.3390/rs10091423. URL: https://www.mdpi.com/2072-4292/10/9/1423.

[39] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[40] Bichen Wu et al. "SqueezeDet: Unified, Small, Low Power Fully Convolutional Neural Networks for Real-Time Object Detection for Autonomous Driving". In: *CoRR* abs/1612.01051 (2016). arXiv: 1612.01051. URL: http://arxiv.org/abs/1612.01051.

[41] Saining Xie et al. "Aggregated Residual Transformations for Deep Neural Networks". In: *arXiv preprint arXiv:1611.05431* (2016).

[42] Siwei Yang et al. "Reducing the feature divergence of RGB and near-infrared images using Switchable Normalization". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2020, pp. 206–211. DOI: 10.1109/CVPRW50498.2020.00031.

[43] Fisher Yu and Vladlen Koltun. "Multi-Scale Context Aggregation by Dilated Convolutions". In: *International Conference on Learning Representations (ICLR)*. May 2016.

[44] Hang Zhang et al. *ResNeSt: Split-Attention Networks*. 2020. arXiv: 2004.08955 [cs.CV].

[45] Michael R. Zhang et al. "Lookahead Optimizer: k steps forward, 1 step back". In: *CoRR* abs/1907.08610 (2019). arXiv: 1907.08610. URL: http://arxiv.org/abs/1907.08610.

[46] Bolei Zhou et al. *Semantic Understanding of Scenes through the ADE20K Dataset*. 2018. arXiv: 1608.05442 [cs.CV].

[47] Jun-Yan Zhu et al. "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks". In: *CoRR* abs/1703.10593 (2017). arXiv: 1703.10593. URL: http://arxiv.org/abs/1703.10593.