Improved Crop and Weed Detection with Diverse Data Ensemble Learning

Muhammad Hamza Asad[1], Saeed Anwar[2,3], and Abdul Bais[1]
[1] University of Regina, Regina SK, Canada
[2] King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia
[3]SDAIA-KFUPM Joint Research Center for Artificial Intelligence, Dhahran, Saudi Arabia
{maq541,abdul.bais}@uregina.ca, saeed.anwar@kfupm.edu.sa

Abstract

Modern agriculture heavily relies on Site-Specific Farm Management practices, necessitating accurate detection, localization, and quantification of crops and weeds in the field, which can be achieved using deep learning techniques. In this regard, crop and weed-specific binary segmentation models have shown promise. However, uncontrolled field conditions limit their performance from one field to the other. To improve semantic model generalization, existing methods augment and synthesize agricultural data to account for uncontrolled field conditions. However, given highly varied field conditions, these methods have limitations. To overcome the challenges of model deterioration in such conditions, we propose utilizing data specific to other crops and weeds for our specific target problem. To achieve this, we propose a novel ensemble framework. Our approach involves utilizing different crop and weed models trained on diverse datasets and employing a teacher-student configuration. By using homogeneous stacking of base models and a trainable meta-architecture to combine their outputs, we achieve significant improvements for Canola crops and Kochia weeds on unseen test data, surpassing the performance of single semantic segmentation models. We identify the UNET meta-architecture as the most effective in this context. Finally, through ablation studies, we demonstrate and validate the effectiveness of our proposed model. We observe that including base models trained on other target crops and weeds can help generalize the model to capture varied field conditions. Lastly, we propose two novel datasets with varied conditions for comparisons. Our code will be available at github.com.

1. Introduction

Throughout human history, farming has been the most crucial component of society for the survival of human beings. Modern farming requires more than traditional techniques and hugely relies on Site Specific Farm Management (SSFM) which requires timely and accurate detection, localization and quantification of crop and weeds in the field [18]. SSFM suggests varying the fertilizer and herbicide application rate to the field [33] by mapping the variability of crops and weeds. With the recent advances in deep learning, accuracy for object detection, localization, and quantification in digital images has tremendously enhanced [48]. Semantic segmentation is widely employed due to its ability to accurately determine the boundaries of different plant categories in an image [12, 45].

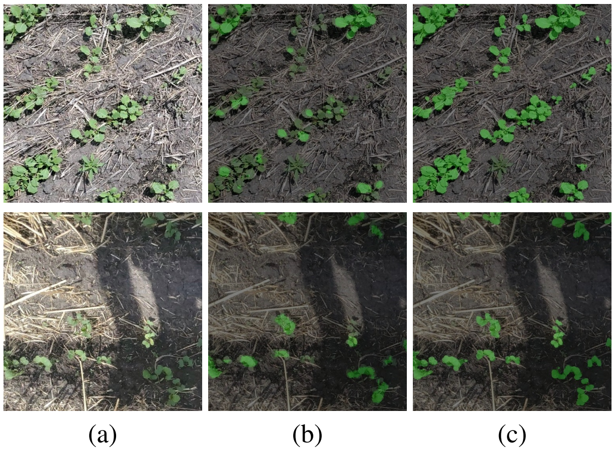

Figure 1. (a) Sample images containing early and mid-stage Canola plants, (b) Canola pixels classified by traditional encode-decoder scheme: ResNet50-SegNet. It can be observed that some Canola plant pixels are misclassified as background class. (c) Our proposed framework addresses the false negatives and rightly classifies the majority of Canola pixels. (Best viewed on screen and when zoomed-in).

Crop and weed-specific binary semantic segmentation models are commonly employed for improved performance and reducing the effort to manually label every single plant category individually at the pixel level [2, 3, 35]. However, with digital images collected under uncontrolled field conditions, the performance of crop and weed-specific semantic segmentation models deteriorates. Varying image backgrounds, crop staging, crop stress, incidence of unseen vegetation, variable ambient lighting conditions and changing parameters of imaging equipment are common causes of model failure from one field to another. Improving models on test data is an active area of research in semantic segmentation-based models [7, 10, 16, 20, 21, 40, 49, 52, 53]. The challenges presented by agriculture imagery collected under uncontrolled field conditions warrant developing solutions for model generalization.

Few works in the literature [28, 30, 32, 36] apply domain adaptation to the semantic segmentation models bridging the domain gap by augmenting labelled data using Generative Adversarial Networks (GANs) [30]. Other methods include adversarial domain adaptation, where a model learns features robust to domain changes [28, 32, 36]. In agriculture applications, image enhancement methods are applied to augment data to generalize semantic segmentation models [38, 43, 54]. However, only some works [35] achieved generalization by learning from the data of other crops and weeds.

We develop a novel method that employs an ensemble framework to achieve generalization as models trained for a different target (crop or weed) task can bridge any domain gap. To avoid the need for end-to-end ensemble prediction, we use teacher-student configuration to train the student model from an ensemble of crop and weed models. In such a setting, two ensemble strategies are used: heterogeneous and homogeneous ensemble [7]. We opt for homogeneous stacking of the base models utilizing different crop models trained on diversified datasets. As each base model is trained on a different crop, fusing the output of the base teacher models is performed using a trainable meta-architecture. Using this methodology, we improve the mean Intersection Over Union (mIOU) for the Canola crop by up to 12% and for Kochia weed by 6% on unseen test data compared to single ResNet50-SegNet semantic segmentation. We also observe that the UNET meta-architecture performs better than other meta-architectures.

We claim the following three contributions.

- We introduce homogeneous stacking of different crop and weed models, which is not investigated before.
- We propose a novel knowledge distillation framework which ensembles different crop/weed teacher models using semantic meta-architecture.
- We evaluate the proposed model by performing different ablation studies.

2. Related Work

Currently, the encoder-decoder framework is widely employed in semantic segmentation for crop and weed detection [22]. UNET is such a network that has symmetrical layers of encoder and decoder. Feature maps from each encoder layer are connected to the corresponding decoder layer [34]. Deep learning architectures like VGG [37] and ResNet50 are used as the encoder block with UNet. Another widely employed encoder-decoder architecture is SegNet [4]. In SegNet, indices of pooling layers are transferred from encoder layers to corresponding decoder upsampling layers. Using these networks, mIOU of 82% is achieved for weed mapping in Canola fields [3]. UNet architectures are tailored for agriculture data to improve crop and weed mapping [41, 54]. These encoder-decoder frameworks' performance decreases with changing scale of the objects and reduced feature resolution. To address these challenges, recently, DeepLab [9] has been used for crop and weed discrimination [18, 50]. Though the object scale changes problem is addressed, these crop and weed detection methods are sensitive to field conditions. With a slight domain shift, the model's performance deteriorates. To bridge domain gaps, augmentation methods are mainly employed in agriculture data [38, 43, 54]. These methods include random image cropping and patching [38, 54], image enhancement techniques [43], and traditional augmenters applied to agriculture images [41]. However, synthetic and augmented data do not account for diverse scenarios of real field conditions.

In deep learning, ensemble methods are widely employed to improve model generalization. However, the ensemble of semantic segmentation models for crop and weed detection is overlooked. To improve model generalization for uncontrolled field conditions, ensemble methods can be adapted to agriculture data. Ensemble learning reduces overall bias and variance by using the collective knowledge of the base models to make predictions. Ensemble learning can be divided into two distinct steps, namely, ensemble strategy and fusion strategy [15]. Next, we discuss these strategies in detail.

2.1. Ensemble Strategies

Ensemble strategies deal with training base models to achieve diversity. Three commonly employed ensemble strategies are 1) bagging, 2) boosting, and 3) stacking.

Bagging methods generate multiple bags of training data to train base models [6]. Bagging in machine learning is employed in techniques like SVM, neural networks and stacked denoising autoencoders [1, 25]. Bagging is helpful in addressing challenging problems such as over-fitting and data imbalance [5, 39].

Boosting Technique: In boosting to make predictions, weak learners are trained with equal weights given to each instance. In subsequent training sessions, more weight is given to the misclassified instances so weak learners can learn from challenging cases. AdaBoost and gradient boosting are commonly employed boosting methods [13, 14].

Recently, boosting has been used with deep neural networks to improve the generalization of the models. Deep belief network, deep boost, incremental boosting CNN, stagewise boosting CNN, and snapshot boosting are found in the literature to improve the effectiveness and efficiency of boosting methods [17, 27, 29, 42, 51].

Stacking strategy: trains a meta-architecture through distinct architectures, algorithms or hyper-parameter settings and combines the base learners' outputs. Welchowski *et al*. [46] and Wang *et al*. [44] improved generalization and reduced bias through different variants of deep convex nets and deep stacking networks, respectively. In medical imaging, Das *et al*. [11] operate different encoder-decoder architectures like SegNet [4] and UNet [34] to classify brain tumours through semantic segmentation. In some cases, features at different stages of the network are also extracted and fused for improved performance of the deep neural network [8].

Apart from the above categorization of ensemble strategies, another categorization is homogeneous and heterogeneous ensemble learning [15]. As the name indicates, the homogeneous ensemble model uses the same base models; however, to create diversity in the prediction of the base models, randomness and uncertainty are added to the training data of each base model. On the other hand, Bagging is an example of homogeneous ensemble learning. Contrary to it, heterogeneous ensemble learning uses multiple architectures as a base model with varying computational costs [26]. The output of these base learners is fused for final prediction. Following subsection details ensemble learning output fusion strategies.

2.2. Fusion Strategies

The final prediction of an ensemble model depends on the approach to fuse the outputs. Multiple strategies can achieve this goal, such as averaging, majority voting, *etc*. Although averaging is simple since it's not adaptive, the outcome usually needs improvement [15]. Similarly, majority voting performs well for shallower networks compared to deep neural networks [23]. Further, a trainable meta-layer is also a commonly applied method for finding the weight of base models in the final output of the ensemble. Stacked generalization and super learner methods are also widely used for regression and classification problems [24, 47].

Our method adapts above mentioned ensemble strategies to semantic segmentation applications in agriculture. We develop a homogeneous stacking ensemble with a trainable meta-architecture on the top to fuse the output of base models trained on diversified targets (crops and weeds).

2.3. Methodology

Recently, a few attempts have been made to propose semantic segmentation models for crops and their different growth stages, as well as models specific to weeds. Nevertheless, these individual models did not perform well due to changes in field conditions, *e.g*., ambient lighting. Also, the occurrence of new unseen vegetation types, background soil, and crop residue may fail the model on new fields. Given the availability of data for different target problems, we investigate if the individual train on specific data can be used to account for varying field conditions. Therefore, a homogeneous stacking ensemble of base models trained on different datasets is proposed. Such an ensemble strategy requires a fusion different than simple averaging. Addressing the generalizing semantic segmentation models using ensemble learning may result in a computationally costly end-to-end ensemble, which warrants training of student models[1] from an ensemble of base models in the teacher. Figure 2 illustrates the flow diagram of the proposed framework.

2.4. ArgMax Baseline Ensemble

ArgMax Baseline Ensemble (ABE) is a rule-based baseline model that compares predictions for a pixel with the predictions of the base models (β). The meta-T model (τ) classifies pixels based on the decision of the base models, which predicts it more confidently. For any pixel (x, y) and i binary models in an ensemble $f_i = \beta_i(x, y)$, class C_τ of the (x, y) is given by the following equation:

$$C_\tau = \max(f_i), \tag{1}$$

where $\max(\cdot)$ is the τ and $i \in 1, 2, \cdots, N$.

2.5. Mask Convolution Ensemble (MCE)

Unlike the baseline teacher model, which uses the ArgMax rule, MCE adds a trainable layer to the combined output of the base models in the ensemble. This layer determines the contribution of each base model's prediction to the pixel class. To achieve this, we utilize a 1×1 convolutional layer h with sigmoid activation σ to reduce the prediction masks of the base models in the ensemble to the desired number of classes. It is assumed that adding a trainable layer on top of the base models will help learn a more complex relationship between the outputs of base models and ground truth.

$$C_\tau = \sigma(h(f_i)). \tag{2}$$

Here, the Meta-T architecture is represented as $\sigma(h(f_i))$.

2.6. Meta-SegNet Ensemble (MSNE)

The third ensemble method stems from using multi-stage CNN for classification and localization. The mask convolution method adds a 1×1 convolution layer over the top

[1] We train each student model individually for the specific data type.

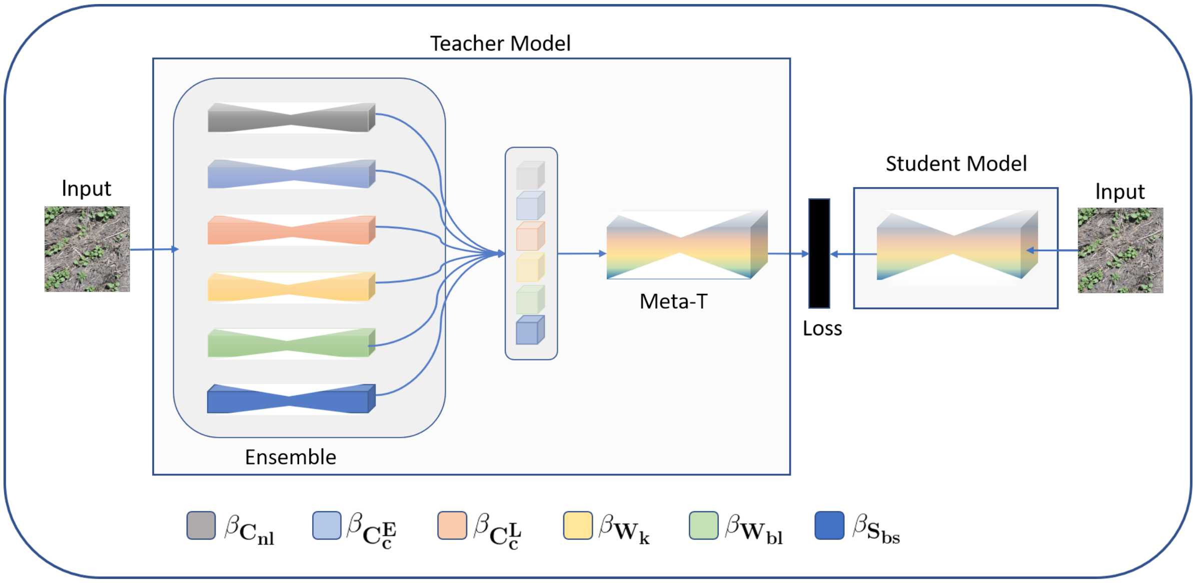

Figure 2. Our proposed framework for the ensemble of the teachers-student model. β's are the base models, which are fused using meta-architecture. Table 1 presents the details of base models and respective datasets used for training. The student model is trained for a specific problem (such as Canola) to learn from the ensemble base models trained on different datasets.

Table 1. Details of base models architecture and respective datasets: Each model is trained on a separate target crop/weed/soil.

Base Models	Pre-trained	Description	Datasets	No. of Images
$\beta_{C_{nl}}$	✓	Detects narrow leaf crops	NLD: Narrow Leaf Dataset	250
$\beta_{C_c^E}$	✓	Detects early stage Canola	ESCCD: Early Stage Canola Crop Dataset	150
$\beta_{C_c^L}$	✓	Detects mid / late stage Canola	LSCCD: Late Stage Canola Crop Dataset	300
β_{W_k}	✓	Detects Kochia weed	KWD: Kochia Weed Dataset	124
$\beta_{W_{bl}}$	✓	Detects broad leaf weeds	BLWD: Early Stage Canola Crop Dataset	150
$\beta_{S_{bs}}$	✓	Detects bare soil	TSD: Total Soil dataset	50

Table 2. The architecture of each proposed model with the ensemble (base models), Meta-T and Student.

	Base		Meta-T		Student	
Models	Enco.	Deco.	Enco.	Deco	.Enco.	Deco.
ABE	R50	Convs	$max(\cdot)$		R50	Convs
MCE	R50	Convs	One Conv		R50	Convs
MSNE	R50	Convs			R50	Convs
MUNE	R50	Convs	R50	Convs	R50	Convs

of base SegNet models, while here, a whole SegNet model ψ is added in the second stage to combine the outputs of base SegNet models in the ensemble. Furthermore, adding a deep network over the top maps the non-linear complex spatial relationships between different objects in the image. While training Meta-SegNet ψ (Meta-T), base model weights are not updated.

$$C_\tau = \psi(f_i) \tag{3}$$

2.7. Meta-UNet Ensemble (MUNE)

Most works in the literature involving ensemble learning combine the output of different architectures trained on the same data [15]. However, the ensemble learning methods used in this study achieve diversification by combining models trained on different datasets. To benefit from the dataset and architectural diversification, we apply UNet as a meta-architecture over the top of SegNet base architecture. We have trained SegNet base models for different crops and weeds. Therefore, the base models' weights are not updated during the training of Meta-UNet ϕ (Meta-T). We hypothesize that using a different architecture for the Meta-T than the base models brings diversity for improved segmentation accuracy.

$$C_\tau = \phi(f_i) \tag{4}$$

2.8. Student Model

The objective is to generalize original base models by sharing the distant learned features of uncontrolled field conditions with each other. After training the ensemble base models in a supervised setting, the student model's Meta-T architecture remains the same as the teacher's Meta-T architecture. The student model is trained unsupervised by minimizing the loss between the teacher and student output.

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N}[C_\tau \log(C_s) + (1 - C_\tau)\log(1 - C_s)]$$

where C_τ and C_s are predictions from teacher and student models, respectively. The architecture of each proposed model for Ensemble, Meta-T and student are shown in Table 2.

3. Experiments

In this section, we provide the details of the experimental setup of our proposed method, followed by our collected datasets. Next, we present detailed comparisons among the state-of-the-art techniques and conclude this section by performing an ablation study, where we analyze the impact of each vital component in the proposed framework.

3.1. Datasets

The proposed methods are tested on two datasets: the Kochia Weed Dataset (KWD) and the Multi-Stage Canola Dataset (MSCD).

Collection Process: High-resolution RGB images are collected from multiple fields using grid sampling for both datasets. Imaging sensors are mounted on farm machinery, collecting ground images in uncontrolled field conditions. Typically, every image contains 3-5 rows of crop. However, sometimes crop rows swell to 10 due to the increased height of imaging equipment while farm machinery turns at the edges of the field. Notably, the data is collected under uncontrolled field conditions. It includes sunny and cloudy conditions and images during dawn and dusk. Some of the images contain farm machinery shadows. Also, sometimes images are blurry due to mechanical vibration. Figure 3 provides insight into the data and variations in the images used for this study.

KWD-2023: Kochia weed infests multiple types of crops, like cereal crops and oilseed crops. Our KWD comes from fields of early stage Canola, late stage Canola, Oats, Wheat and Durum. The dataset consists of 99 images.

MSCD-2023: In comparison, Canola data comes from multiple fields, consisting of two stages: early-stage dicots Canola and mid-stage broad leaves Canola. MSCD-2023 contains 305 images acquired in 2023.

However, it is to be noted that base models in the ensemble of the teacher model are trained on separate and distinct datasets. These comprise cereal crops (Oats, Wheat, Barley and Duram), early and mid-stage Canola, broad-leaf weeds, and Kochia weed.

Table 3. Comparing the performance of Kochia-specific deep learning ensemble models. Boldface shows the best-performing model on the test set for the specific metric

Models	fwIOU	mIOU	IOU Non-Kochia	IOU Kochia
β_{W_k}	0.9256	0.8373	0.9549	0.7198
ABE	0.7331	0.5769	0.7523	0.3686
MCE	0.9278	0.7563	0.9573	0.5553
MSNE	0.9357	0.8614	0.9604	0.7625
MUNE	**0.9371**	**0.8638**	**0.9615**	**0.7638**

3.2. Settings

Backbones: Our proposed framework uses ResNet50 [19] as the backbone encoder block. Furthermore, SegNet [4] architecture is adopted for pre-trained base models in the ensemble. The selection of SegNet in the ensemble models is based on previous works where it performs marginally better than UNet on agriculture data [2, 3, 31]. Moreover, the architecture of the student network is the same as the base models, *i.e.*, SegNet.

Setup: Categorical cross-entropy is the loss function in our proposed framework. The dataset split is 15%-15%-70% for testing, validation and training. The proposed framework is trained with GPU RTX 3090 support. Adam is used as an optimizer with a learning rate of 0.001, batch size of 2 and input image dimensions of 1440 × 1088 × 3. The training dataset is augmented using standard augmenters to avoid overfitting.

Metrics: Classwise IOU, mean IOU and frequency weight IOU (fwIOU) are the performance metrics. Notably, pre-trained base models in the teacher are trained on different datasets particular to the respective target problem - crop/weed. We assume the diverse data of base models will capture different scenarios of uncontrolled field conditions.

3.3. Comparisons

In this case study, we train the earlier mentioned four learning models: ABE, MCE, MSNE and MUNE. For both Kochia and Canola-specific ensemble models, we use six base models trained on different datasets and target problems (crop/weed). Our proposed MCE, MSNE and MUNE models are trained and compared with ABE ensemble and base model β_{W_k}.

Comparisons on KWD-2023: Table 3 presents the results of ensemble methods on KWD-2023. After analyzing the results, we can observe that the MUNE outperforms all other models on all metrics showing slightly higher IOUs than the MSNE. In terms of mIOU, Kochia-specific MUNE performs *2.5%* better than β_{W_k}. Additionally, Kochia IOU demonstrates a *4.5%* enhancement compared to β_{W_k}. We also notice that the learnable ensemble meta-architectures perform significantly better than the baseline ABE. Fur-

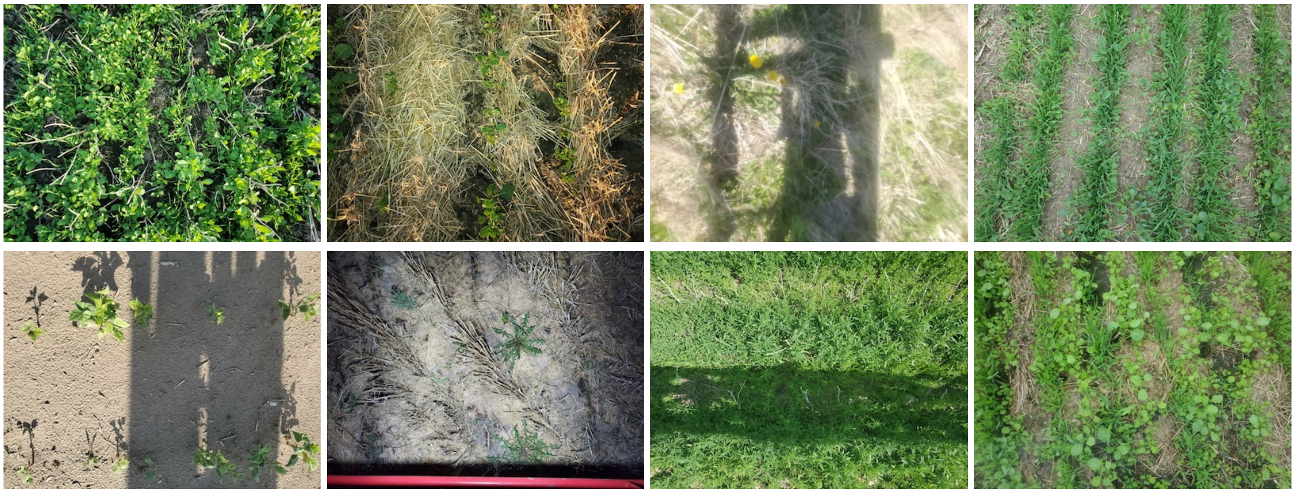

Figure 3. The real world challenging field conditions: the data insight and the challenges posed by uncontrolled field conditions like blurring, variable orientations of crop rows, changing ambient lighting conditions, equipment shadows and images collected during night time under auxiliary lights.

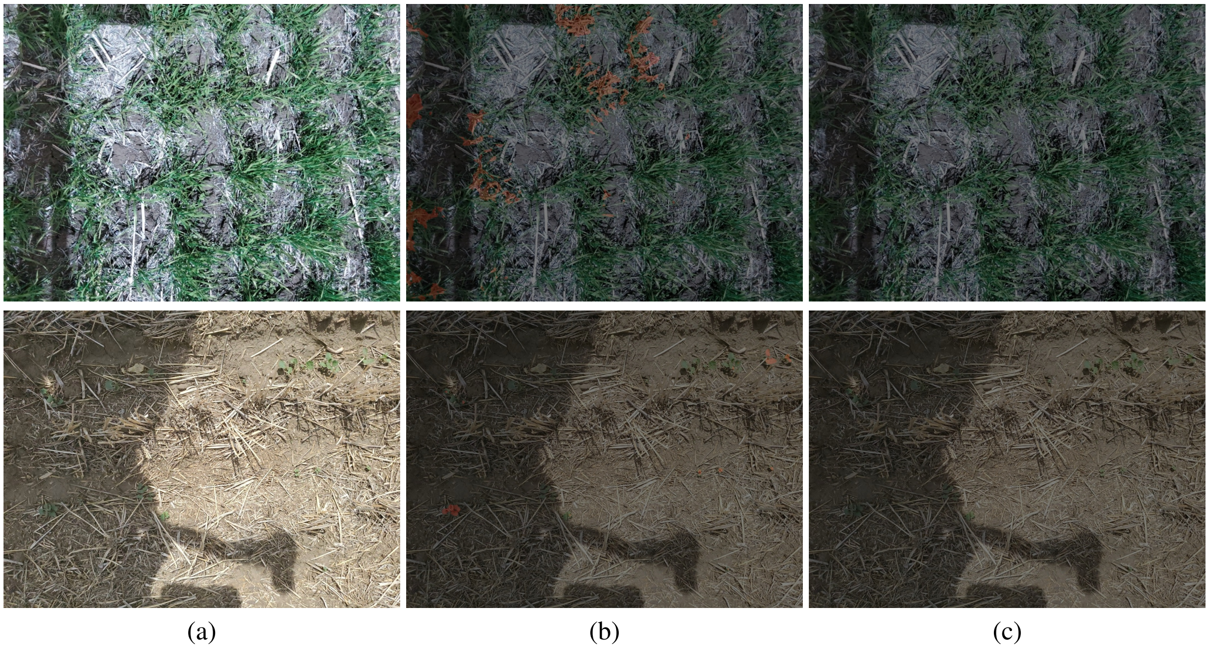

Figure 4. The visual comparisons from β_{W_k} and MUNE models. a) The Groundtruth images, b) The prediction of the β_{W_k} while c) predictions of MUNE on images. The β_{W_k} model detects early-stage Canola and narrow leaf as Kochia (false positives), whereas the ensemble model addresses this problem and removes false positives.

thermore, multi-class segmentation using one-vs-all binary segmentation models is worse than the original one-vs-all models. Based on the Kochia ensemble models, we can infer that if multi-class labels are available, trainable meta-architectures can combine one-vs-all binary segmentation models for multi-class semantic segmentation with improved IOUs. To illustrate the improvements made by ensemble methods over the β_{W_k}, Figure 4 shows some examples and further demonstrate that our proposed model effectively addresses false positive detection of Kochia in Canola and narrow leaf crops via a combination of pre-trained base models. The narrow leaf and Canola models can successfully and confidently detect their respective plants of interest, which helps remove false positive detec-

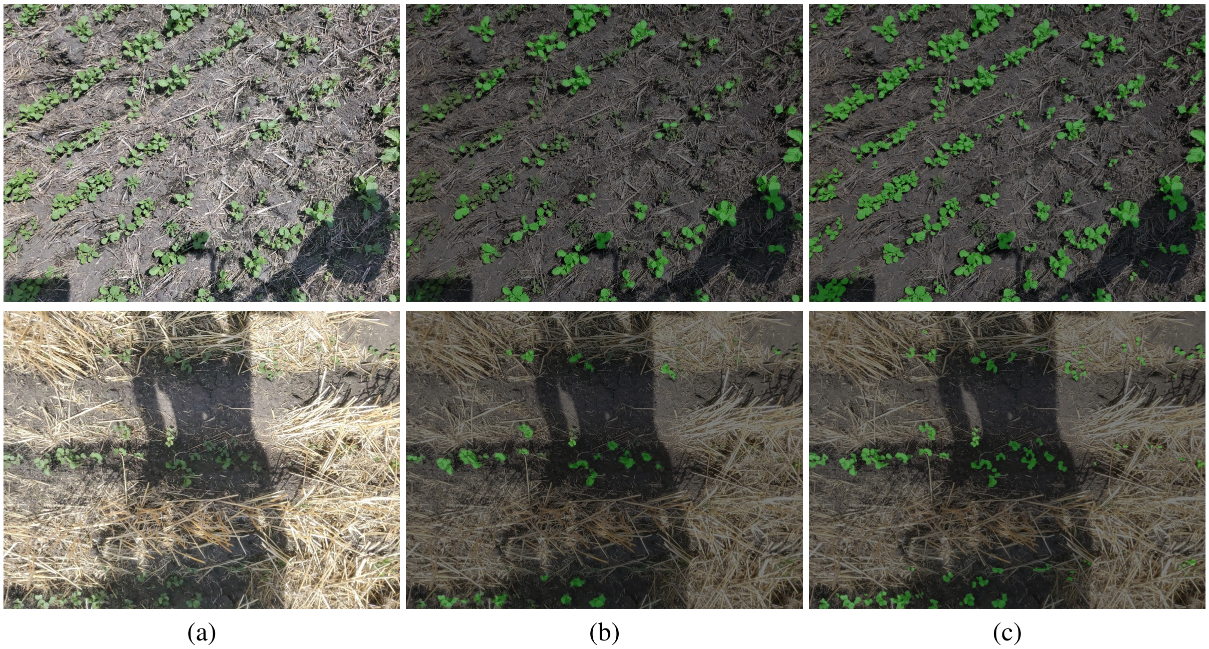

Figure 5. (a) Ground truth, (b) The individual models: $\beta_{C_c^E}$ & $\beta_{C_c^L}$ misses some Canola plants in both early and late stages of the crop (c) our proposed MUNE framework detects missing Canola plants in highly varied field conditions.

Table 4. Comparing the performance of Canola-specific deep learning ensemble models. Boldface shows the best-performing model on the test set for the specific metric.

Models	fwIOU	mIOU	IOU Non-Canola	IOU Canola
FCN_32	0.9587	0.7136	0.9779	0.4493
PSPNet	0.9669	0.7727	0.9821	0.5632
UNet	0.9804	0.8628	0.9895	0.7360
SegNet	0.9766	0.8411	0.9872	0.6950
DeepLab V3+	0.9807	0.8763	0.9858	0.7668
HRNet	0.9832	0.8771	0.9897	0.7643
SegFormer	0.9822	0.8835	0.9802	0.7868
ABE (Ours)	0.9406	0.8093	0.9510	0.6676
MCE (Ours)	0.9838	0.8881	0.9913	0.7848
MSNE (Ours)	0.9811	0.8698	0.9899	0.7497
MUNE (Ours)	**0.9859**	**0.9040**	**0.9923**	**0.8159**

tion of Kochia in the ensemble Kochia settings. Table 4 demonstrates the ensemble model's effectiveness in detecting all Canola stages by building upon four base models: $\beta_{C_{nl}}$, $\beta_{C_c^E}$, $\beta_{C_c^L}$, and β_{W_k}.

Comparisons on MSCD-2023: Table 4 summarises the results of our proposed framework on MSCD-2023. Like KWD 2023, the MUNE model shows the best performance among different ensemble methods with mIOU improvement of 9% from the ABE model and 5% from $\beta_{C_c^E}$ & $\beta_{C_c^L}$ combined. If we compare class-wise Canola IOU, it improves by 14.8% as compared to the ABE and 12% from combined Canola models. Figure 5 presents examples highlighting performance improvements made by our proposed framework. The prediction masks in Figure 5(b) are predicted through $\beta_{C_c^E}$ & $\beta_{C_c^L}$ model, while prediction masks on the right are predicted through the Meta-UNet ensemble models. It can be observed that under predictions are made by the original model, as some Canola plants are altogether missed. At the same time, ensemble MUNE significantly addresses the mentioned false negative problem and detects Canola plants to the edges of the crop leaves.

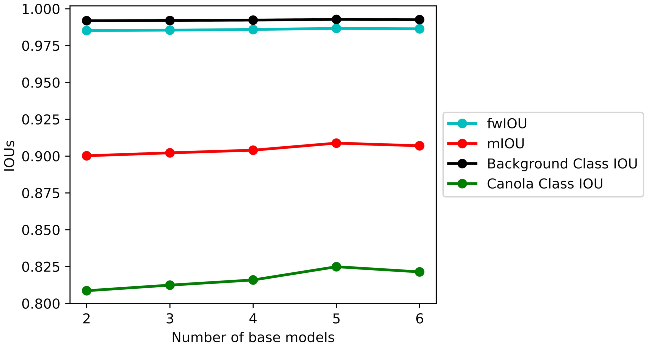

Figure 6. IOUs Vs. the number of base models: It can be observed that increasing the number of base models brings diversification in the ensemble, improving IOUs.

3.4. Ablation Studies

In this paper, we perform two types of ablation studies. The first study compares the IOU metrics of MUNE model

Table 5. Comparing the IOU variations with respect to changing the number of base models. The best results are achieved when five base models are used in the ensemble of teachers with the MUNE model.

Models	$\beta_{C_c^E}$	$\beta_{C_c^L}$	β_{W_k}	$\beta_{C_{nl}}$	$\beta_{S_{bs}}$	$\beta_{W_{bl}}$	fwIOU	mIOU	IOU Non-Canola	IOU Canola
$M1$	✓	✓					0.9852	0.9002	0.9919	0.8086
$M2$	✓	✓	✓				0.9855	0.9022	0.9920	0.8124
$M3$	✓	✓	✓	✓			0.9859	0.9040	0.9923	0.8159
$M4$	✓	✓	✓	✓	✓		**0.9867**	**0.9088**	**0.9928**	**0.8249**
$M5$	✓	✓	✓	✓	✓	✓	0.9864	0.9070	0.9926	0.8214

Table 6. Comparing the inference time, Floating Point Operations (FLOPs) and total parameters for end-to-end ensembles.

Models	$\beta_{C_c^E}$	$\beta_{C_c^L}$	β_{W_k}	$\beta_{C_{nl}}$	$\beta_{S_{bs}}$	$\beta_{W_{bl}}$	Inference Time	FLOPs	Parameters
FCN_32							0.80	1.28 T	451 M
PSPNet							0.39	0.204 T	29 M
UNet							0.50	0.572 T	16 M
SegNet							0.48	0.428 T	14 M
DeepLab V3+							-	0.320T	24 M
$E1$	✓	✓					0.50	1.86 T	62 M
$E2$	✓	✓	✓				0.54	2.3 T	77 M
$E3$	✓	✓	✓	✓			0.57	2.7 T	92 M
$E4$	✓	✓	✓	✓	✓		0.61	3.1 T	107 M
$E5$	✓	✓	✓	✓	✓	✓	0.73	3.6 T	122 M

by varying the number of base models. The second study presents changes in efficiency for end-to-end ensembles with the change in the number of base models. We restrict these ablation studies to MSCD. Figure 6 shows the trend of IOU metrics with respect to the increasing number of base models. We take only those combinations of base models whose target crop/weed is abundant in the field. For example, we included $\beta_{C_{nl}}$ as the base model because it detects narrow leaf crops as well as narrow leaf weeds in the images. It can be observed that IOUs improve with the inclusion of more base models. The peak of IOUs comes when five base models are included in the ensemble of the teacher model. The other reason could be the inclusion of $\beta_{S_{bs}}$ (bare soil extractor) in the ensemble, which helps the model capture different soil backgrounds in the images under variable lighting conditions. However, with the inclusion of the broad leaf base model ($\beta_{W_{bl}}$), the performance drops. It may be due to the confusion caused by $\beta_{W_{bl}}$ with $\beta_{C_c^L}$ as both models learn features of broad leaves. Notably, a plant type could be a crop in one field, but its occurrence in other fields could be deemed as a weed. Table 5 summarizes the results of the first ablation study. Similarly, Table 6 presents the inference time, FLOPs and parameters of the end-to-end ensembles. It can be observed that inference time for E1 to E5 models changes from 0.50 seconds to 0.73 seconds, and FLOPs changes from 1.86 T to 3.6 T. Deteriorating efficiency of end-to-end ensemble warrants training student models to improve inference time and require less computational resources.

4. Conclusion

In this paper, we present an ensemble of base models for teacher framework to improve the performance of semantic segmentation models for crop and weed under uncontrolled field conditions. Existing methods attempt to achieve model generalization through augmentation and agriculture data synthesis. However, these methods struggle to capture numerous scenarios of uncontrolled field conditions. To address these challenges, we propose a teacher model trained on diversified target crops and weeds to teach a student model for our target crop/weed. In addition, in the teacher model, we propose a meta-architecture to fuse the outputs of base models trained on different target problems to enhance semantic segmentation performance for crop and weed detection. Our framework will pave the way for research in the cross-applicability of different crop and weed-specific models to

Acknowledgement

This research project is supported by an NSERC Alliance - Mitacs Accelerate grant (NSERC Ref: ALLRP 581076 - 22 & Mitacs Ref: IT34971), titled ”Developing Machine Learning Methods for RGB Images to Quantify Crop and Weed Populations Across Agricultural Fields” in partnership with Croptimistic Technology Inc.

References

[1] Ricardo F Alvear-Sandoval and Aníbal R Figueiras-Vidal. On building ensembles of stacked denoising auto-encoding classifiers and their further improvement. *Information Fusion*, 39:41–52, 2018. 2

[2] Muhammad Hamza Asad and Abdul Bais. Weed density estimation using semantic segmentation. In *Image and Video Technology: PSIVT 2019 International Workshops, Sydney, NSW, Australia, November 18–22, 2019, Revised Selected Papers 9*, pages 162–171. Springer, 2020. 1, 5

[3] Muhammad Hamza Asad and Abdul Bais. Weed detection in canola fields using maximum likelihood classification and deep convolutional neural network. *Information Processing in Agriculture*, 7(4):535–545, 2020. 1, 2, 5

[4] 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(12):2481–2495, 2017. 2, 3, 5

[5] Jerzy Błaszczyński and Jerzy Stefanowski. Neighbourhood sampling in bagging for imbalanced data. *Neurocomputing*, 150:529–542, 2015. 2

[6] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. 2

[7] Chen-Hao Chao, Bo-Wun Cheng, and Chun-Yi Lee. Rethinking ensemble-distillation for semantic segmentation based unsupervised domain adaption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2610–2620, 2021. 2

[8] Li Chen, Xin Dou, Jian Peng, Wenbo Li, Bingyu Sun, and Haifeng Li. EFCNet: Ensemble full convolutional network for semantic segmentation of high-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5, 2021. 3

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE PAMI*, 40(4):834–848, 2017. 2

[10] Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6830–6840, 2019. 2

[11] Suchismita Das, Srijib Bose, Gopal Krishna Nayak, and Sanjay Saxena. Deep learning-based ensemble model for brain tumor segmentation using multi-parametric mr scans. *Open Computer Science*, 12(1):211–226, 2022. 3

[12] LG Divyanth, Aanis Ahmad, and Dharmendra Saraswat. A two-stage deep-learning based segmentation model for crop disease quantification based on corn field imagery. *Smart Agricultural Technology*, 3:100108, 2023. 1

[13] Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *icml*, pages 148–156. Citeseer, 1996. 2

[14] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001. 2

[15] M.A. Ganaie, Minghui Hu, A.K. Malik, M. Tanveer, and P.N. Suganthan. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151, 2022. 2, 3, 4

[16] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2477–2486, 2019. 2

[17] Shizhong Han, Zibo Meng, Ahmed-Shehab Khan, and Yan Tong. Incremental boosting convolutional neural network for facial action unit recognition. *Advances in Neural Information Processing Systems*, 29, 2016. 3

[18] Leila Hashemi-Beni, Asmamaw Gebrehiwot, Ali Karimoddini, Abolghasem Shahbazi, and Freda Dorbu. Deep convolutional neural networks for weeds and crops discrimination from uas imagery. *Frontiers in Remote Sensing*, 3:755939, 2022. 1, 2

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[20] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016. 2

[21] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018. 2

[22] Kun Hu, Zhiyong Wang, Guy Coleman, Asher Bender, Tingting Yao, Shan Zeng, Dezhen Song, Arnold Schumann, and Michael Walsh. Deep learning techniques for in-crop weed identification: A review. *arXiv preprint arXiv:2103.14872*, 2021. 2

[23] Cheng Ju, Aurélien Bibaut, and Mark van der Laan. The relative performance of ensemble methods with deep convolutional neural networks for image classification. *Journal of Applied Statistics*, 45(15):2800–2818, 2018. 3

[24] Cheng Ju, Mary Combs, Samuel D Lendle, Jessica M Franklin, Richard Wyss, Sebastian Schneeweiss, and Mark J van der Laan. Propensity score prediction for electronic healthcare databases using super learner and high-dimensional propensity score methods. *Journal of Applied Statistics*, 46(12):2216–2236, 2019. 3

[25] AS Khwaja, Muhammad Naeem, A Anpalagan, A Venetsanopoulos, and B Venkatesh. Improved short-term load forecasting using bagged neural networks. *Electric Power Systems Research*, 125:109–115, 2015. 2

[26] Zeynep Hilal Kilimci and Selim Akyokuş. Deep learning- and word embedding-based heterogeneous classifier ensembles for text classification. *Complexity*, 2018. 3

[27] Vitaly Kuznetsov, Mehryar Mohri, and Umar Syed. Multi-class deep boosting. *Advances in Neural Information Processing Systems*, 27, 2014. 3

[28] Geun-Ho Kwak and No-Wook Park. Unsupervised domain adaptation with adversarial self-training for crop classification using remote sensing images. *Remote Sensing*, 14(18): 4639, 2022. 2

[29] Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong. Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1805–1812, 2014. 3

[30] Yuzhen Lu, Dong Chen, Ebenezer Olaniyi, and Yanbo Huang. Generative adversarial networks (gans) for image augmentation in agriculture: A systematic review. *Computers and Electronics in Agriculture*, 200:107208, 2022. 2

[31] Xu Ma, Xiangwu Deng, Long Qi, Yu Jiang, Hongwei Li, Yuwei Wang, and Xupo Xing. Fully convolutional network for rice seedling and weed image segmentation at the seedling stage in paddy fields. *PloS one*, 14(4):e0215676, 2019. 5

[32] Yuchi Ma and Zhou Zhang. Multi-source unsupervised domain adaptation on corn yield prediction. In *AI for Agriculture and Food Systems*, 2022. 2

[33] David Mulla and Raj Khosla. Historical evolution and recent advances in precision farming. *Soil-specific farming precision agriculture*, pages 1–35, 2016. 1

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2, 3

[35] Bishwa B Sapkota, Chengsong Hu, and Muthukumar V Bagavathiannan. Evaluating cross-applicability of weed detection models across different crops in similar production environments. *Frontiers in Plant Science*, 13:837726, 2022. 1, 2

[36] Aleksandr Yu Shkanaev, Dmitry L Sholomov, and Dmitry P Nikolaev. Unsupervised domain adaptation for dnn-based automated harvesting. In *Twelfth International Conference on Machine Vision (ICMV 2019)*, pages 243–249. SPIE, 2020. 2

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[38] Daobilige Su, He Kong, Yongliang Qiao, and Salah Sukkarieh. Data augmentation for deep learning based semantic segmentation and crop-weed classification in agricultural robotics. *Computers and Electronics in Agriculture*, 190:106418, 2021. 2

[39] Dacheng Tao, Xiaoou Tang, Xuelong Li, and Xindong Wu. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(7):1088–1099, 2006. 2

[40] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1379–1389, 2021. 2

[41] Hafiz Sami Ullah, Muhammad Hamza Asad, and Abdul Bais. End to end segmentation of canola field images using dilated U-Net. *IEEE Access*, 9:59741–59753, 2021. 2

[42] Elad Walach and Lior Wolf. Learning to count with CNN boosting. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 660–676. Springer, 2016. 3

[43] Aichen Wang, Yifei Xu, Xinhua Wei, and Bingbo Cui. Semantic segmentation of crop and weed using an encoder-decoder network and image enhancement method under uncontrolled outdoor illumination. *Ieee Access*, 8:81724–81734, 2020. 2

[44] Bin Wang, Bing Xue, and Mengjie Zhang. Particle swarm optimisation for evolving deep neural networks for image classification by evolving and stacking transferable blocks. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE, 2020. 3

[45] Dashuai Wang, Wujing Cao, Fan Zhang, Zhuolin Li, Sheng Xu, and Xinyu Wu. A review of deep learning in multiscale agricultural sensing. *Remote Sensing*, 14(3):559, 2022. 1

[46] Thomas Welchowski and Matthias Schmid. A framework for parameter estimation and model selection in kernel deep stacking networks. *Artificial Intelligence in Medicine*, 70: 31–40, 2016. 3

[47] David H Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. 3

[48] Xiongwei Wu, Doyen Sahoo, and Steven CH Hoi. Recent advances in deep learning for object detection. *Neurocomputing*, 396:39–64, 2020. 1

[49] Zuxuan Wu, Xintong Han, Yen-Liang Lin, Mustafa Gokhan Uzunbas, Tom Goldstein, Ser Nam Lim, and Larry S Davis. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 518–534, 2018. 2

[50] Helong Yu, Minghang Che, Han Yu, and Jian Zhang. Development of weed detection method in soybean fields utilizing improved deeplabv3+ platform. *Agronomy*, 12(11): 2889, 2022. 2

[51] Wentao Zhang, Jiawei Jiang, Yingxia Shao, and Bin Cui. Snapshot boosting: a fast ensemble framework for deep neural networks. *Science China Information Sciences*, 63:1–12, 2020. 3

[52] Zhedong Zheng and Yi Yang. Unsupervised scene adaptation with memory regularization in vivo. *arXiv preprint arXiv:1912.11164*, 2019. 2

[53] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision*, 129(4):1106–1120, 2021. 2

[54] Kunlin Zou, Xin Chen, Yonglin Wang, Chunlong Zhang, and Fan Zhang. A modified u-net with a specific data argumentation method for semantic segmentation of weed images in the field. *Computers and Electronics in Agriculture*, 187: 106242, 2021. 2