Precision Agriculture (2023) 24:2002–2018
https://doi.org/10.1007/s11119-023-10027-7

A W-shaped convolutional network for robust crop and weed classification in agriculture

Syed Imran Moazzam[1,2] · **Tahir Nawaz**[1,2] · **Waqar S. Qureshi**[1,3] · **Umar S. Khan**[1,2] · **Mohsin Islam Tiwana**[1,2]

Accepted: 24 April 2023

Abstract

Agricultural image and vision computing are significantly different from other object classification-based methods because two base classes in agriculture, crops and weeds, have many common traits. Efficient crop, weeds, and soil classification are required to perform autonomous (spraying, harvesting, etc.) activities in agricultural fields. In a three-class (crop–weed–background) agricultural classification scenario, it is usually easier to accurately classify the background class than the crop and weed classes because the background class appears significantly different feature-wise than the crop and weed classes. However, robustly distinguishing between the crop and weed classes is challenging because their appearance features generally look very similar. To address this problem, we propose a framework based on a convolutional W-shaped network with two encoder–decoder structures of different sizes. The first encoder–decoder structure differentiates between background and vegetation (crop and weed), and the second encoder–decoder structure learns discriminating features to classify crop and weed classes efficiently. The proposed W network is generalizable for different crop types. The effectiveness of the proposed network is demonstrated on two crop datasets—a tobacco dataset and a sesame dataset, both collected in this study and made available publicly online for use by the community—by evaluating and comparing the performance with existing related methods. The proposed method consistently outperforms existing related methods on both datasets.

Keywords Crops and weeds · Pixel-level classification · Semantic segmentation · Weed detection

---

✉ Waqar S. Qureshi
waqar.qureshi@tudublin.ie

[1] Department of Mechatronics Engineering, National University of Sciences and Technology, H-12, Islamabad, Pakistan

[2] Robot Design and Development Lab, National Centre of Robotics and Automation (NCRA), National University of Sciences and Technology, H-12, Islamabad, Pakistan

[3] School of Computer Science, Technological University Dublin, Dublin, Ireland

Introduction

Efficient crop, weeds, and soil classification are prerequisites for autonomous spraying, harvesting, crop health monitoring, and weeding activities in agricultural fields (Hashemi-Beni et al., 2022; Milioto et al., 2018; Subeesh et al., 2022). Semantic segmentation (Sa et al., 2018) offers a solution based on the classification of prediction of every pixel into three classes. As reported in the literature (You et al., 2020), the classification of background is generally achieved with a high accuracy, which is significantly different feature-wise from the vegetation (crop and weed classes); however, the difficulty lies in distinguishing between crop and weed classes that have resemblance in colour and leaf structure. Secondly, in some previous works, the background is shown to be effectively removed using linear thresholding methods, as done in Ferreira et al. (2017). Still, this strategy poses a challenge for accurately classifying vegetation, particularly in variable lighting conditions. The top background subtraction techniques currently utilised are based on deep neural networks and have significantly improved performance in contrast with traditional unsupervised approaches (Bouwmans et al., 2019). So, deep neural network-based background removal is desirable, as they are expected to handle non-linear lighting conditions better.

To address the challenges mentioned above, we propose a deep learning framework based on a convolutional W-shaped network (a W network). The key innovation point of the proposed network is the usage of optimised two encoder–decoder structures connected in series for achieving the desired pixel-level classification as opposed to traditional approaches based on a single encoder–decoder configuration. The first encoder–decoder structure differentiates between background and vegetation. The second encoder–decoder structure primarily aims to learn discriminating crop and weed features in the background-removed images, which has been found to robustly and efficiently classify vegetation (crop and weed) classes in this study. As a part of the experimental validation, we show that the proposed framework is generalisable to multiple crop types. This has been demonstrated by training the proposed W network from scratch on our collected tobacco crop dataset and then fine-tuning for our collected sesame crop dataset using transfer learning. The proposed method shows encouraging results for both crop types compared to the existing related methods. We have made both datasets available online (Moazzam, 2023) for the research community.

Related work

Efficient crop, weed, and soil classification is critical for autonomous agricultural activities such as spraying and harvesting. Traditional thresholding-based methods for background removal suffer from limitations in variable lighting conditions, leading to the removal of vegetation pixels. To accurately classify vegetation and background pixels, there is a need for a learning-based method. Previous studies have attempted to address this issue using techniques such as Otsu adaptive thresholding, normalization, ExG-ExR indices, histogram equalization, and morphological operations. However, these methods still have limitations in terms of coarse background removal, inappropriate contrast enhancement, and difficulty in selecting appropriate thresholds for real-world aerial images. Therefore, the need for an improved learning-based method for background removal is necessary, making suitable deep learning techniques desirable. Milioto et al. (2017) and Espejo-Garcia et al. (2020) used Otsu's thresholding method on NDVI and normalized RGB channels, respectively, but both faced difficulties in variable lighting conditions. Knoll et al. (2019) used the HSV color space, while Le et al. (2020) used ExG-ExR indices, and Jiang et al. (2019) used histogram equalization, but all faced limitations in background removal. Alam et al. (2020) utilized morphological operations to distinguish between soil and vegetation, but this approach changed the image data at the corners, making the background removal coarse. Therefore, the use of suitable deep learning techniques is crucial for improving background removal in agricultural image and vision computing.

Crop/weed classification using classical machine learning-based methods

Classical machine learning-based methods have been employed for crop and weed classification; Sabzi et al. (2020) utilized 13 color features, 8 shape features, 8 texture features, and 5 moment-invariant features, whereas Karimi et al. (2006) and Wendel and Underwood (2016) employed SVM and LDA to classify plants. However, as feature engineering remains a challenge in classical machine learning, the application of deep learning is preferred to extract thousands of features automatically. Additionally, Ishak et al. (2007) suggested that the neural network-based technique can be improved by adding convolutional layers to capture more discriminative features. Therefore, deep learning-based methods are preferable for larger datasets as they can extract more discriminative features automatically.

Object detection-based deep learning methods for crop/weed classification

In recent years, object detection-based deep learning methods have been used for crop/weed classification, relying on vegetation blob or bounding box detection within an image (Nkemelu et al., 2018; Partel et al., 2019). These methods, of course, require reliable bounding box annotations and image-level annotations. This category of methods is generally computationally efficient but has limitations in terms of localisation when weeds are in close proximity or are occluded by crops. This category of methods make use of Faster RCNN (Jiang et al., 2020) and YOLO family neural networks like YOLO-v3 (Sharpe et al., 2020), YOLOv4 (Zhao et al., 2022), YOLOv5 (Wang et al., 2022), YOLOv6 (Dang et al., 2023), and YOLOv7 (Gallo et al., 2023). These deep learning neural networks are efficient, however, there are two major problems found in their implementation. The first problem is mixed detection and bounding boxes overlap. This way, crop and weed detection become ambiguous as rectangular boxes could contain both classes. The second problem found in the implementation of YOLO models is missed detections for small weeds. Therefore, pixel-wise deep learning application is recommended if a fine outline of crop and weed plants is required.

Semantic segmentation-based methods for crop/weed classification

This sections highlights significant pixel-level crop–weed classification methods. These deep learning models provide inference for every pixel in the image, this way the resultant detections of crop, weed and background show a smooth profile outline. Sa et al. (2018) and Abdalla et al. (2019) proposed semantic segmentation-based frameworks involving pixel-level classification of crops and weeds in the field. Kamath et al. (2022) applied

SegNet and UNet to classify weeds in paddy crops. Hashemi-Beni et al. (2022) also used SegNet and UNet for crop and weed classification in a sugarcane dataset. However, these networks did not demonstrate encouraging results in terms of the classification of crops and weeds. This is apparently due to the usage of a single classifier to distinguish among three classes (background, crop, and weed); as the background class is more distinctive, it gets classified more accurately, whereas crop and weed classes need more attention. Better pixel accuracy is expected to be achieved using a sequential concatenation of two semantic segmentation models as proposed by Kim and Park (2022), MTS-CNN network composed of two UNet models connected as one. It has two stages, each using an encoder size of four. This is the best semantic segmentation classifier so far to the best of our knowledge however, there remains room to improve further and optimise this network to differentiate between background, crop, and weed efficiently.

Semantic segmentation is the best fit deep learning technique if we want pixel-level classification or if we want a fine profile of detections of crop, weed and background classes. Moreover, a two stage application of semantic segmentation is more accurate as compared to single stage semantic segmentation in the case where target is weed classification in agricultural crops as suggested by Kim and Park (2022). Kim and Park (2022) proposed a two-stage semantic segmentation model that shows promising results for crop and weed classification, and we propose to optimize this model further.

To optimize two-stage semantic segmentation-based framework, we propose a simpler model to distinguish between background and vegetation in the first stage and a model with more neurons to then (expectedly) better distinguish between crop and weed classes in the second stage. To this end, we proposed a deep learning-based fully convolutional W-shaped network that uses two encoder–decoder structures with variable encoder sizes coupled in series to achieve better pixel-level classification.

Datasets

We collected a tobacco and a sesame crop dataset for this study. The tobacco dataset is captured using Mavic mini drone gimbal camera 1/2.3″ CMOS sensor in Mardan, Khyber Pakhtunkhwa, Pakistan, and the sesame dataset is captured using an Agrocam NDVI sensor in Ballo Shahabal Village near Jhang, Punjab, Pakistan. These datasets are captured in early stage of the crops, different fields of tobacco crop dataset are captured after 15 to 40 days after emergence (DAE) of plants and different fields of sesame crop dataset are captured after 16 to 45 DAE of plants. The image capture resolution of both datasets is 1 920 × 1 080 pixels. Due to hardware and software limitations, the images are divided into non-overlapping patches of size 480 × 352 pixels, which are then used for training and testing. The tobacco dataset is captured at an average altitude of 4 m and the sesame dataset is captured at an average altitude of 4.5 m, corresponding to ground sampling distance (GSD) of 0.1 cm/pixel and 0.3 cm/pixel, respectively. We conducted eight fly campaigns for each dataset. For training we have selected the fields which have more diversity in them in terms of different weeds and different sizes of weeds, this helps in term of better training of neural networks. Testing is done very extensively in our research by choosing multiple fields other than the fields on which training is done, this practice helps in achievement of better generalizable model. For testing of tobacco crop, total images used in training and testing are 864 and 1 656 respectively, which shows train/test percentage of 35:65, however the readers should not be confused by bigger percentage of testing data, as this complete testing data belongs to seven different tobacco fields. Similarly in the case of sesame crop, total images used in training and testing are 12 00 and 720 respectively, which shows train/test percentage of 62:38, and the testing data belongs to six different sesame fields. The MATLAB Image Labeller app is used to label both datasets. The tobacco dataset offers RGB imagery, whereas the sesame dataset offers NGB imagery (i.e., NIR, Green, and Blue channels). The data capturing campaign for both of tobacco and sesame datasets are shown in Fig. 1 with respect to time, date and number of images used in experiments. The soil and sunlight conditions in these datasets are not quantifiable. There is a variability in the soil and sunlight that makes classification of these datasets more challenging.

Proposed network architecture

Shortcomings in the existing pixel-wise classification methods

When compared to the soil background class, the semantic segmentation of crop and weed classes performs poorly, which is one of its flaws, as we have seen in the literature. In the past, when semantic segmentation has been used, soil pixels were very accurately classified, whereas the performance of crop and weed classification was inferior as the crop and weed pixels were confused between each other. This led to the increase in the number of false positives and false negatives, thereby lowering the accuracy of weed classification, as reported in the quantitative results in earlier works (Abdalla et al., 2019; Kamath et al., 2022). The more likely explanation for this issue is that background classification is frequently accurate because it differs greatly from vegetation (classes of crop and weed), in terms of features. However, it might be difficult to discern between crop and weed classifications since their colours and leaf structures are similar. We recommend the use of two encoder–decoder structures that are paired in sequence to enhance crop and weed pixel-wise categorisation. The first encoder–decoder structure could identify between background and vegetation on background-removed pictures,

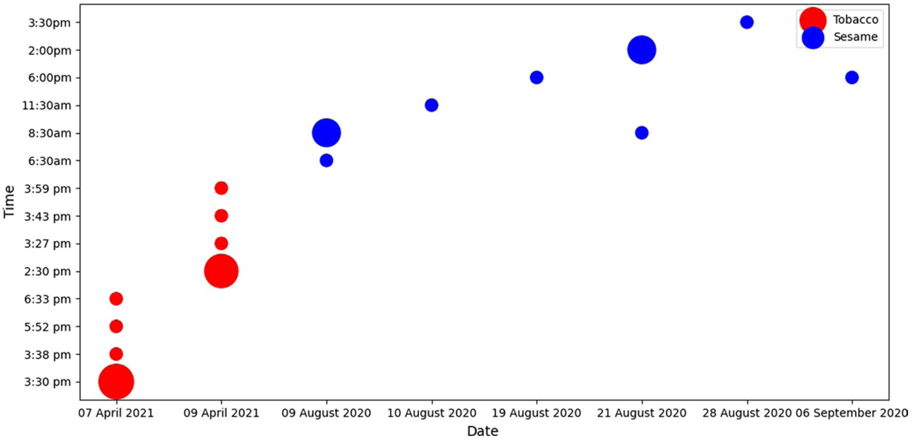


Fig. 1 The bigger red circle is 936 on 7th April and 864 for 9th April. The small red circle represents 120 images. Similarly for sesame, the small blue circle represents 120 images, and the bigger blue represents 600 images. The tobacco data is taken on consecutive days; however, the sesame data is spread over a period of 2 months of plant growth (Color figure online)

whereas the second encoder–decoder structure could learn to distinguish between crop and weed traits.

Innovation point of proposed w network

As opposed to the conventional method of semantic segmentation, which employs just one encoder–decoder structure for pixel-wise classification of all classes, our proposed W network uses two encoder–decoder structures for pixel-wise classification. The second encoder–decoder structure in our proposed W network has a unique job to do, and that is to learn better aspects of both kinds of vegetation, such as crops and weed, which are difficult to tell apart because of their close similarities.

There is a scientific basis for the suggested W network; for example, vegetation and background are extremely distinct groups that can be clearly distinguished from one another, unlike crop and weed classifications, which have many qualities in common. Furthermore, we noticed that crop and weed classification is less accurate than background classification in the literature, which is how we came up with the concept of two encoder–decoder structures.

Proposed W network

Our proposed W network (Moazzam et al., 2023) takes three-channel image input. The W network has two encoder–decoder structures. The first encoder–decoder structure is responsible for differentiating between vegetation and background, and it has an encoder size of two. We selected the encoder size by experimenting with an encoder size of two, three, and four and chose the encoder size of two that maximizes the classification performance, keeping computational complexity to a minimum.

After the first encoder–decoder structure, we added a background removal layer before the second encoder–decoder structure. This layer removes background pixels, and these images without background are fed into the second encoder–decoder structure. The second structure is trained separately from the first structure on crop, weed, and background classes, and then it is added to the first structure after the background removal layer. The second structure learns discriminative crop and weed features better when the background-removed images are used for training. It has an encoder size of three, again chosen experimentally by varying encoder sizes. Crop and weed have a higher appearance similarity than background and vegetation, and that's why a higher encoder size is required here compared to the first structure.

The encoder–decoder structures in the proposed W network could incorporate different backbone and segmentation networks. We experimented with SegNet and UNet segmentation networks and used Vanilla, Vanilla Mini, VGG16, MobileNet, and ResNet50 as backbones. As part of training and testing, we used non-overlapping patches of the size 224×224 for MobileNet as this is the maximum size it operates on and the size 480×352 for the remaining ones. Our experimentation showed UNet as the best segmentation network. As a backbone to UNet, Vanilla Mini and VGG16 showed the best pixel classification results in the first and second encoder–ecoder structures, respectively. A detailed architecture of the proposed W network is shown in Fig. 2.

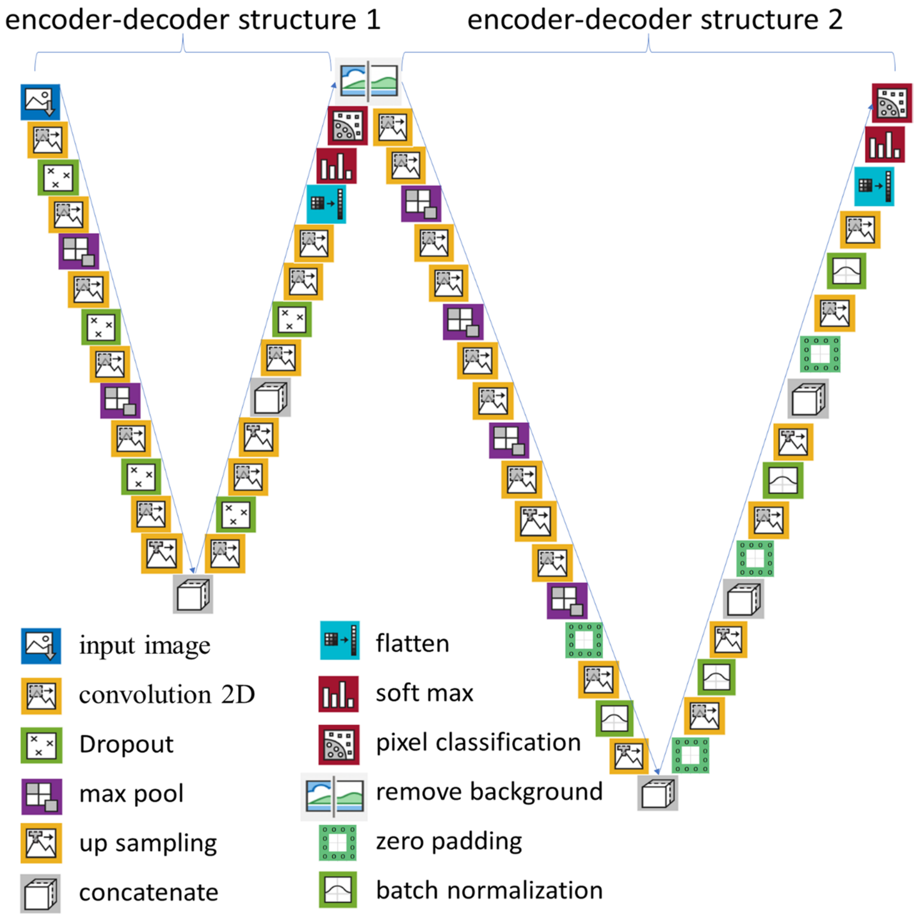


Fig. 2 Architecture of the proposed W network. The structure 1 has an encoder size of two and the structure 2 has an encoder size of three. The arrows show a flow of data in layers sequentially. Every square box in the neural network shows a different layer within network

Computational complexity of proposed W network

Here in this section, we analyze the trainable and untrainable parameters of the proposed W network's encoder–decoder structures in comparison to UNet. Deep neural networks' complexity is demonstrated by these parameters. The proposed W network's first encoder–decoder structure contains 471586 trainable parameters and 0 untrainable ones, whereas the second encoder–decoder structure has 12321603 trainable parameters and 1920 untrainable ones. In comparison to the proposed W network, UNet has roughly 12321603 trainable parameters utilizing the VGG16 backbone. Overall, the proposed W network has more computational complexity than UNet even when using the same backbone, and this complexity increase is due to the incorporation of additional 471586 trainable parameters in the proposed W network's initial encoder–decoder structures.

Implementation of proposed W network

Data augmentation of vertical and horizontal flips are applied, which are expected to result in better model learning. Binary cross-entropy and categorical cross-entropy are chosen as the loss function in 1st and 2nd encoder–decoder structures, and Adam is selected as the optimizer. There is usage of both vertical and horizontal data augmentation. The epoch with the smallest validation loss is used to save the best-trained model. Furthermore, as for the number of parameters, the W network has expectedly got a larger number of parameters due to its two encoder–decoder structures as compared to UNet and SegNet. For example, with VGG16 backbone, for W network the first encoder–decoder structure has 471586 trainable and zero non-trainable parameters, and the second encoder–decoder structure has 12321603 trainable and 1920 non-trainable parameters. On the other hand, UNet and SegNet both have approximately 12 million parameters with this backbone model. We used the mean intersection over union (MIOU), pixel accuracy, and F1-score as the evaluation metrics.

Results and discussion

Figure 3 shows the performance of the proposed W network on different tobacco fields in terms of MIOU. The tobacco dataset has seven test fields, having different soil and sunlight conditions. The proposed W network consistently shows encouraging performance.

Note that the comparatively smaller MIOU on campaign number 3 is likely because the data contains early-stage minuscule weed (difficult to be seen even with a naked eye) that could be missed by the method. One important thing to mention here is normally one dataset is divided into training and testing, which gives higher classification results, opposite to that completely different datasets, which are acquired in different fields conditions are used in testing results shown in Fig. 3.

In most of the researches we saw in literature, a normal practice is to divide the data in train–test split, this practice generate higher accuracy results as both train and test data is taken under same conditions, however if we use completely separate datasets which are taken at different location with different timing and lighting conditions, for training and testing then it is a challenging situation. We have experimented with both

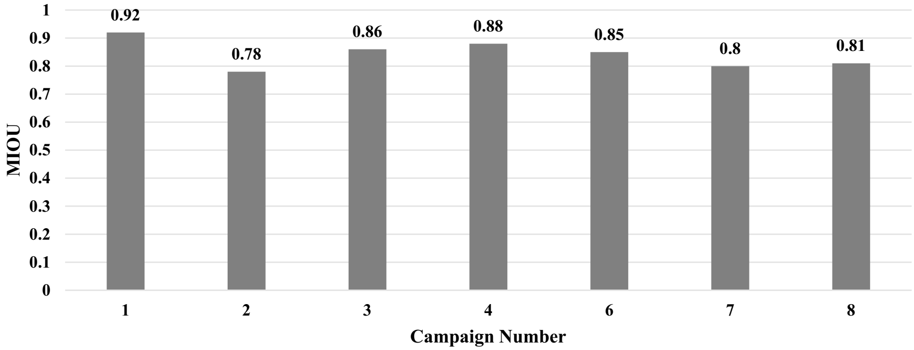


Fig. 3 MIOU of the W network on tobacco fields. Note that the campaign number 5 is used for training

same field training and testing and separate field training and testing to show that our trained models are not overfitting or underfitted, the slightly low MIOU in the case of separate field training and testing are due to different field conditions.

Using the same field number ‘2’ for training and testing with a 70/30 split, and we discovered that the MIOU was significantly higher as compared to when field number ‘1’ was used for training and field number ‘2’ for testing. Using the same field for training and testing will always produce better results than using separate fields for training and testing. Figure 4 compares the outcomes for these two different train–test configurations. While it is true that using the same field for training and testing can sometimes produce better results than using separate fields, it is not a universal truth. Our intention was to suggest that there can be benefits to using the same field for training and testing in certain cases, particularly when dealing with small datasets or when there is a lack of diversity in the available fields. However, we acknowledge that there are potential drawbacks to using the same field for training and testing, including overfitting and lack of generalization to other fields as shown in Fig. 4.

The reason behind selection of UNet and SegNet is that they are extensively used in recent related articles, e.g. Sa et al. (2018), Abdalla et al. (2019), Kamath et al., (2022), Hashemi-Beni et al. (2022) and Kim and Park (2022), where these networks are used to solve agriculture crop–weed classification problem with encouraging performance.

Regarding the selection of Vanilla, Vanilla Mini, and MobileNet as backbones, the reason is their adaptability and computational performance for real-time application. These networks provide lower computational complexity and faster inference when used as a backbone. We thought it would be useful to show the effectiveness of these lighter-weight models in comparison to more computationally heavy models, which is why we also experimented with the well-known VGG16 and ResNet50 models.

We have selected different variation of UNet and SegNet as benchmark to validate the proposed W network. We have conducted extensive experiments to validate our proposed W network. Table 1 highlights six different experiments under same conditions of proposed W network and the selected benchmark which showed superiority of proposed W network as depicted by results shown in Figs. 5, 6 and 7.

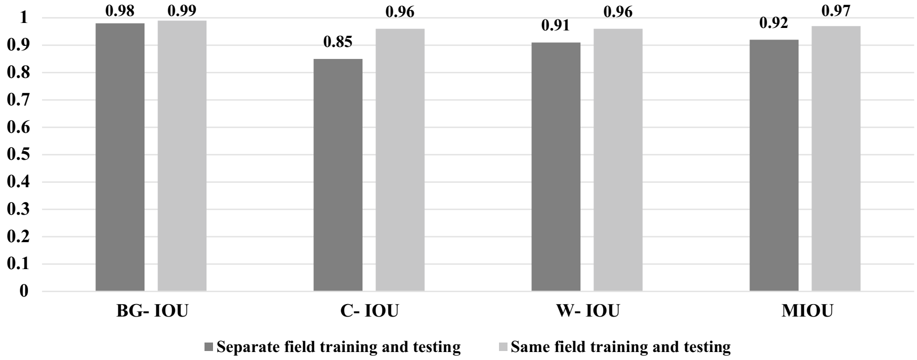


Fig. 4 Accuracy difference with separate and same field training and testing (*BG* background, *C* crop, *W* weed, *IOU* intersection over Union, *MIOU* mean intersection over Union)

Table 1 Comparisons of Proposed W Network with Benchmark semantic segmentation models

Experiment No	Proposed W Network	Benchmark	Dataset
1	W network with VGG16 backbone	UNet with VGG16 backbone	Tobacco
2	W network with Vanilla Mini backbone	UNet with Vanilla Mini backbone	Tobacco
3	W network with MobileNet backbone	UNet with MobileNet backbone	Tobacco
4	W network with Vanilla backbone	SegNet with Vanilla backbone	Tobacco
5	W network with ResNet50 backbone	SegNet with ResNet50 backbone	Tobacco
6	Finetuned W Network	UNet	Sesame

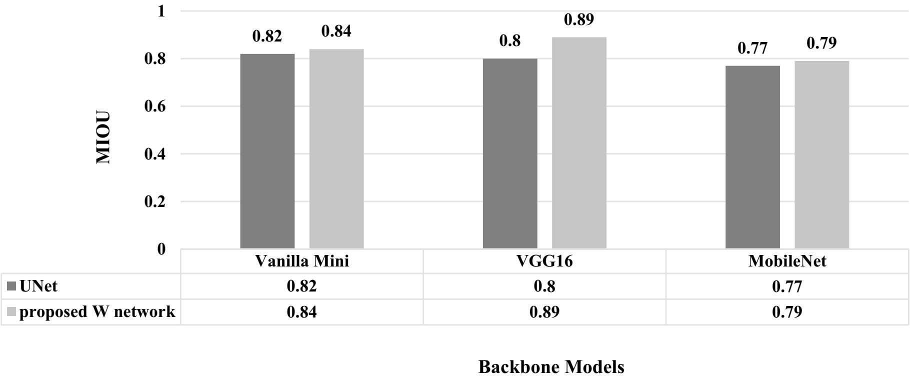


Fig. 5 Comparison of the UNet and W network with different backbone models on tobacco dataset

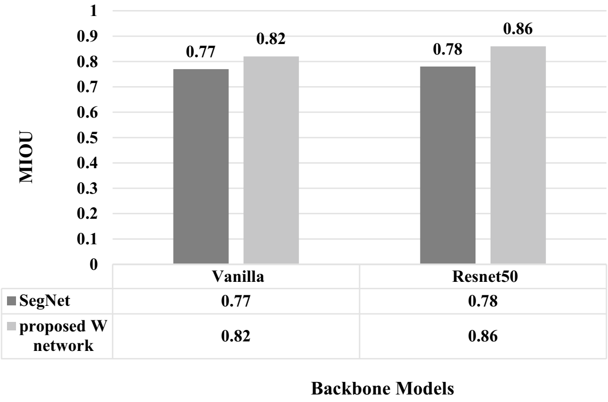
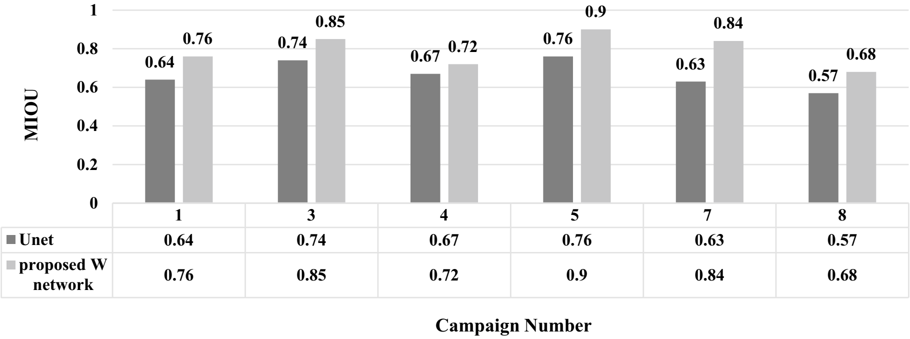


Fig. 6 Comparison of the SegNet and W network with different backbone models on tobacco dataset

Figures 5 and 6 compare UNet and SegNet with the W network using the tobacco dataset from all test campaigns with different backbones within these networks. The results show that the proposed W network consistently outperforms UNet and SegNet.



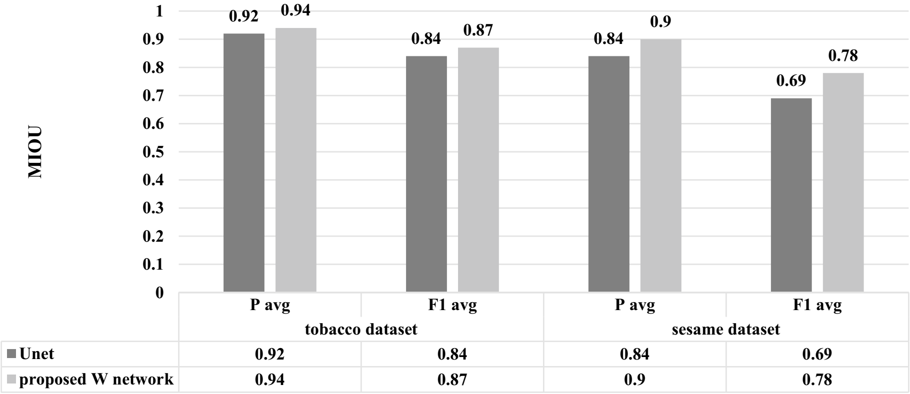
	1	3	4	5	7	8
■ Unet	0.64	0.74	0.67	0.76	0.63	0.57
■ proposed W network	0.76	0.85	0.72	0.9	0.84	0.68

Fig. 7 Comparison of the UNet and fine-tuned W network on different fields of sesame crop. Campaign numbers 2 and 6 are used for fine-tuning tobacco-trained model

To test the generalization ability of the proposed W network, we fine-tuned and adapted the trained network on the tobacco dataset for the sesame dataset. All the layers of the W network are fine-tuned except the background removal layer, using 1 200 images of sesame for 50 epochs.

We evaluated the effectiveness of the proposed W network on images from six different sesame fields and compared the performance against UNet with Vanilla Mini backbone (Fig. 7) as this combination showed the best performance among all of the backbone combinations with UNet and SegNet (Figs. 5, 6). The results show that the W network performs better than UNet (Fig. 7).

For a more holistic evaluation, we also show a performance comparison of the proposed W network with UNet based on the pixel accuracy (P) and F1-score measures both on tobacco and sesame datasets. Figure 8 shows the cumulative performance in terms of the average accuracy (Pavg) and average F1-score (F1avg) computed by averaging the corresponding values across all test images. The proposed W network consistently outperforms UNet on both tobacco and sesame datasets.



	P avg	F1 avg	P avg	F1 avg
	tobacco dataset		sesame dataset	
■ Unet	0.92	0.84	0.84	0.69
■ proposed W network	0.94	0.87	0.9	0.78

Fig. 8 Comparison of the UNet and W network on both datasets based on average accuracy (Pavg) and average F1-score (F1avg)

We have compared our proposed W network with (Kim & Park, 2022) MTS-CNN two-stage network, which consists of two UNets connected in series. The input image sizes for this experiment were fixed at 480×352 for both MTS-CNN network and our proposed W network, which is helpful for comparing the outcomes. In our experiment, we keep encoder–decoder sizes for MTS-CNN at three for both UNet stages. In the proposed W network, the first stage employs an encoder size of two, and the second stage an encoder size three. The implementation hyperparameters for our suggested modal and the MTS-CNN are kept the same. A comparison of the proposed model with MTS-CNN is shown in Fig. 9.

Although the outcomes from the two approaches are comparable, the proposed W network is more computationally efficient. The complexity of the MTS-CNN network increased by use of the same size encoder in both phases. Therefore, we advised utilizing UNet with encoder sizes 2 and 3 for the two stages respectively in our proposed method.

As we can see, our suggested model (471 586 + 12 321 603 = 12 793 189 trainable parameters) is considerably less computationally complex than MTS-CNN (12 321 603 + 12 321 603 = 24 643 206 trainable parameters) when comparing computational complexity of both models using same 480×352 size input images. As a result, the proposed W network model we've developed can be seen as an optimized version of MTS-CNN with just around half the computing complexity.

Figure 10 compares the proposed W network with UNet on key images from the tobacco dataset. Likewise, Fig. 11 compares the proposed W network with UNet on key images from the sesame dataset. The green bounding boxes show some key areas of interest. The W network performed better than UNet on both tobacco and sesame datasets, as the pixel-level classification of the W network is more accurate than the UNet as shown in Figs. 10 and 11.

We can see in Fig. 10 that our suggested W network application of semantic segmentation improves the categorisation and separability of classes for tobacco and weed. We can see difficult lighting circumstances in three of the Fig. 10 photos, with direct sunlight in some places and shade in others. Effective weed and tobacco detection in these pictures demonstrates how resistant to changing lighting conditions our suggested method is. With green rectangular boxes, we've highlighted significant weed locations in Figs. 10 and 11 where our suggested approach has demonstrated higher class separability.

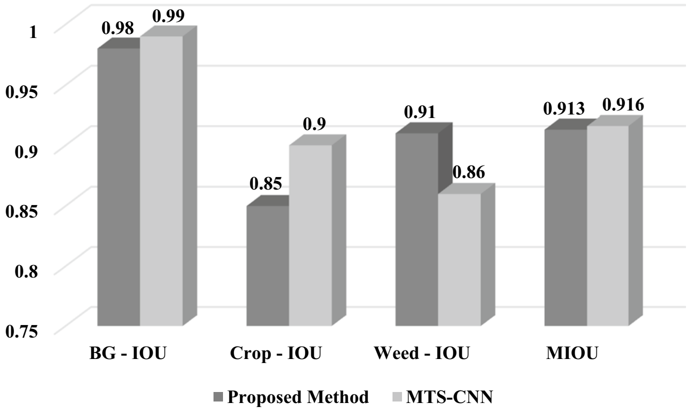
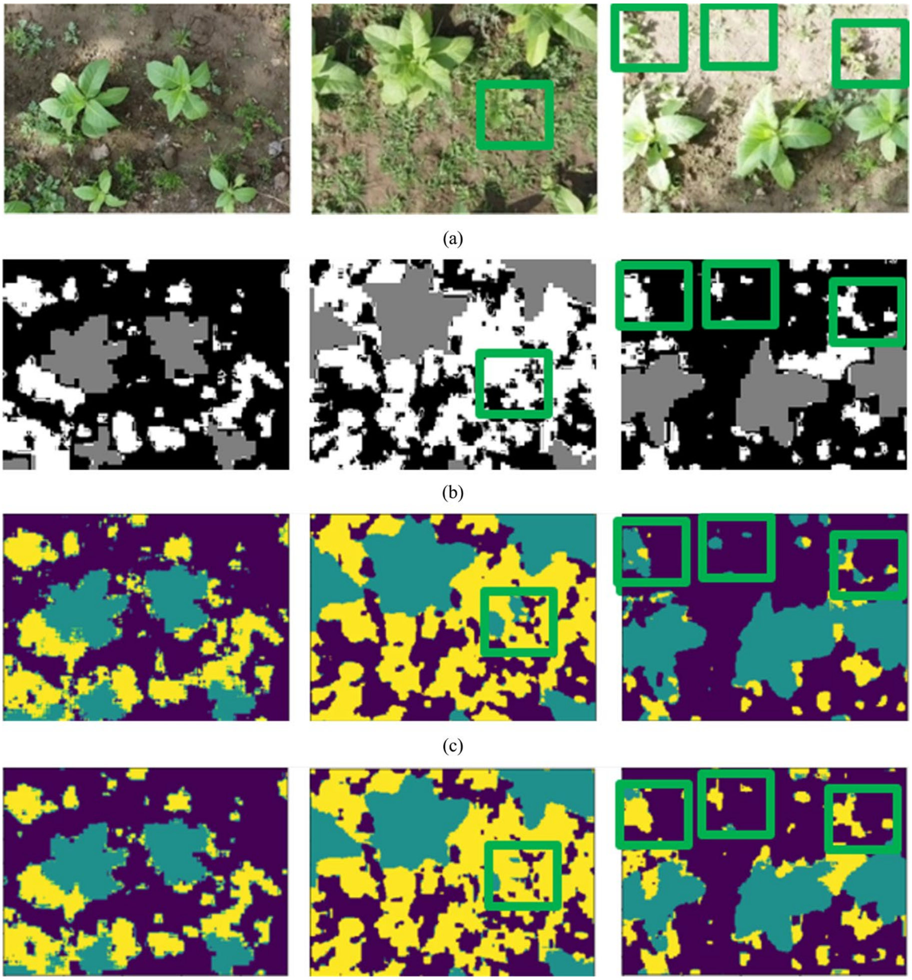


Fig. 9 Comparison of the proposed model with MTS-CNN



(a)

(b)

(c)

(d)

Fig. 10 Qualitative comparison of UNet and the proposed W network on tobacco dataset. In predicted results, yellow color represents weed pixels, cyan color represents pixels classified as tobacco, and dark blue color represents background. **a** Three key test images from tobacco dataset with different soil and sunlight conditions, **b** corresponding ground truth of the images in **a**, and **c**, d predicted results using UNet (**c**) and the proposed W network (**d**). Green rectangular boxes show better crop/weed prediction using the W network (Color figure online)

Conclusions

We proposed a new deep learning-based approach for classifying crops, weeds, and backgrounds in agricultural field applications. The proposed method is based on training a W-shaped network consisting of two encoder–decoder structures: the first structure removes the background, and the second structure learns discriminative features for classifying crops and weeds. We showed the effectiveness of the proposed W network by evaluating and comparing the performance with related state-of-the-art semantic

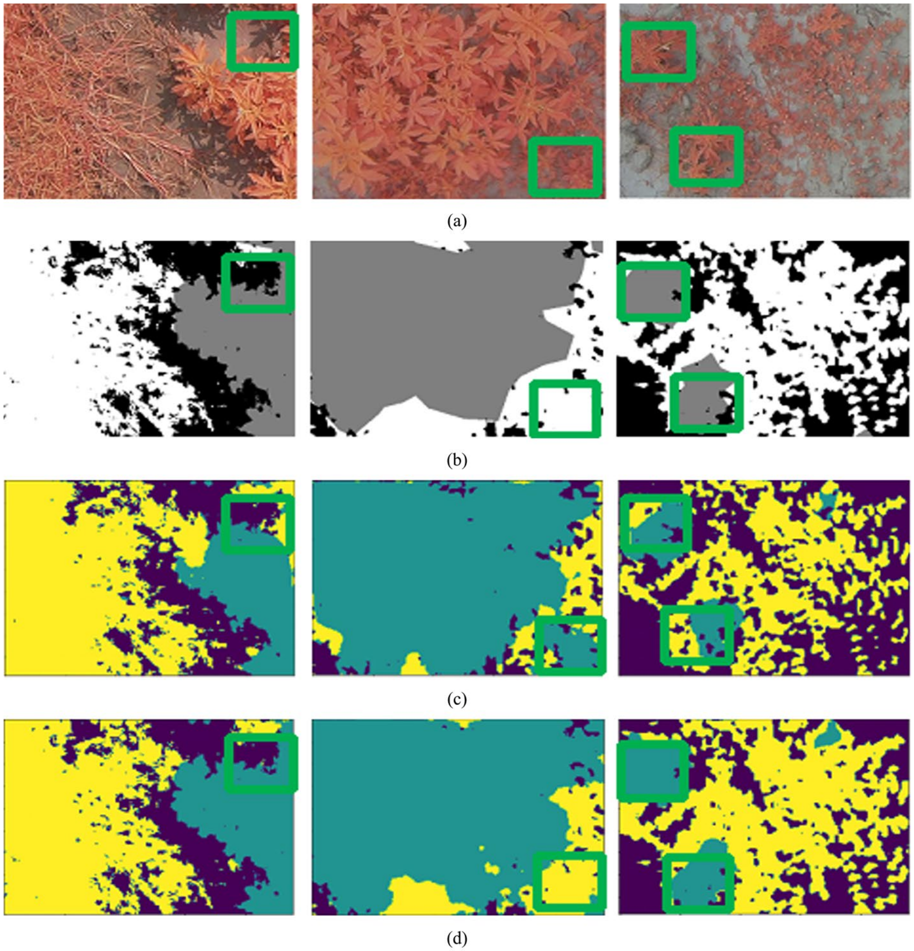


Fig. 11 Qualitative comparison of UNet and the proposed W network on sesame dataset. In predicted results, yellow color represents weed pixels, cyan color represents pixels classified as sesame, and dark blue color represents background. **a** Three key test images from sesame dataset with different soil and sunlight conditions, **b** corresponding ground truth of the images in **a**, and **c**, **d** predicted results using UNet (**c**) and the proposed W network (**d**). Green rectangular boxes show better sesame/weed prediction using W network (Color figure online)

segmentation networks, i.e., SegNet and UNet, on two new aerial agricultural datasets (a new RGB tobacco dataset and a new NGB sesame crop dataset). We collected the dataset as a part of this study and made it publicly available online for the community. In the tobacco dataset, the W network is trained directly from scratch. In contrast, on the sesame dataset, we showed the adaptability of the proposed W network by fine-tuning it with the tobacco learned network. Due to this adaptability of proposed W network, it could be finetuned for other similar crops. To the best of our knowledge, no other researcher have done weed detection from aerial images in tobacco and sesame crops, so this research sets the benchmark and provides first customized solution of weed classification in aerial images for tobacco and sesame crops.

The results showed that the proposed W network outperformed existing related approaches (SegNet and UNet) under the same neural network backbone models on same datasets. Indeed, the experimental evidence shows that the W network is equally effective whether used directly (i.e., learning from scratch for a particular crop type) or indirectly (transfer learning and fine-tuning for a different crop type) for background–crop–weed classification applications. In future work, we aim to test further the proposed W network's generalisation capability on other crop types and heights from the crop. The limitation of the study is that it only focuses on two specific crops, tobacco and sesame, and thus, the generalizability of the proposed W network to other crop types at different imaging heights is not adequately tested. A potential application of this study could involve autonomous aerial spraying of agrochemicals on tobacco and sesame crops to treat weeds, pests, insects, and diseases. Also, an accurate application of chemicals is expected to reduce soil pollution and address health-related concerns.

Funding Open Access funding provided by the IReL Consortium. This research is supported by the Higher Education Commission of Pakistan and the National Centre of Robotics and Automation under Grant Number DF 1009-0031.

Data availability The data supporting this study's findings and analyzed during the current study are available from the corresponding author upon reasonable request. The image dataset is available at Moazzam, Imran (2023). Tobacco Dataset https://data.mendeley.com/datasets/5dpc5gbgpz, Sesame Dataset https://data.mendeley.com/datasets/9pgv3ktk33. *Mendeley*.

Declarations

Conflict of interest The authors, S. I. Moazzam, U. S. Khan, T. Nawaz, W. S. Qureshi, and Mohsin Islam Tiwana have no conflicts of interest to disclose.

References

Abdalla, A., Cen, H., Wan, L., Rashid, R., Weng, H., Zhou, W., & He, Y. (2019). Fine-tuning convolutional neural network with transfer learning for semantic segmentation of ground-level oilseed rape images in a field with high weed pressure. *Computers and Electronics in Agriculture, 167*, 105091.

Alam, M., Alam, M. S., Roman, M., Tufail, M., Khan M. U., & Khan, M. T. (2020). Real-time machine-learning based crop/weed detection and classification for variable-rate spraying in precision agriculture. In *Seventh international conference on electrical and electronics engineering (ICEEE)*, Antalya, Turkey, 2020 (pp. 273–280). https://doi.org/10.1109/ICEEE49618.2020.9102505.

Bouwmans, T., Javed, S., Sultana, M., & Jung, S. K. (2019). Deep neural network concepts for background subtraction: A systematic review and comparative evaluation. *Neural Networks*. https://doi.org/10.48550/arXiv.1811.05255

Dang, F., Chen, D., Lu, Y., & Li, Z. (2023). YOLOWeeds: A novel benchmark of YOLO object detectors for weed detection in cotton production systems. *Computers and Electronics in Agriculture, 205*, 107655.

Espejo-Garcia, B., Mylonas, N., Athanasakos, L., Fountas, S., & Vasilakoglou, I. (2020). Towards weeds identification assistance through transfer learning. *Computers and Electronics in Agriculture, 171*, 105306. https://doi.org/10.1016/j.compag.2020.105306

Ferreira, A. D. S., Freitas, D. M., Silva, G. G. D., Pistori, H., & Folhes, M. (2017). Weed detection in soybean crops using CONVnets. *Computers and Electronics in Agriculture, 143*, 314–324. https://doi.org/10.1016/j.compag.2017.10.027

Gallo, I., Rehman, A. U., Dehkordi, R. H., Landro, N., La Grassa, R., & Boschetti, M. (2023). Deep object detection of crop weeds: Performance of YOLOv7 on a real case dataset from UAV images. *Remote Sensing, 15*(2), 539. https://doi.org/10.3390/rs15020539

Hashemi-Beni, L., Asmamaw, G., Ali, K., Abolghasem, S., & Freda, D. (2022). Deep convolutional neural networks for weeds and crops discrimination from UAS imagery. *Frontiers in Remote Sensing*. https://doi.org/10.3389/frsen.2022.755939

Ishak, A. J., Mokri, S. S., Mustafa, M. M., & Hussain, A. (2007). Weed detection utilizing quadratic polynomial and ROI techniques. In *Fifth student conference on research and development*, Selangor, Malaysia, 2007 (pp. 1–5). https://doi.org/10.1109/SCORED.2007.4451360.

Jiang, H., Zhang, C., Qiao, Y., Zhang, Z., Zhang, W., & Song, C. (2020). CNN feature based graph convolutional network for weed and crop recognition in smart farming. *Computers and Electronics in Agriculture., 174*, 105450. https://doi.org/10.1016/j.compag.2020

Jiang, Y., Li, C., & Paterson, A. H. (2019). DeepSeedling: Deep convolutional network and Kalman filter for plant seedling detection and counting in the field. *Plant Methods, 15*, 141. https://doi.org/10.1186/s13007-019-0528-3

Kamath, R., Balachandra, M., Vardhan, A., & Maheshwari, U. (2022). Classification of paddy crop and weeds using semantic segmentation. *Cogent Engineering, 9*, 1. https://doi.org/10.1080/23311916.2021.2018791

Karimi, Y., Prasher, S. O., Patel, R. M., & Kim, S. H. (2006). Application of support vector machine technology for weed and nitrogen stress detection in corn. *Computers and Electronics in Agriculture, 51*(1–2), 99–109. https://doi.org/10.1016/j.compag.2005.12.001

Kim, Y. H., & Park, K. R. (2022). MTS-CNN: Multi-task semantic segmentation-convolutional neural network for detecting crops and weeds. *Computers and Electronics in Agriculture, 199*, 107146. https://doi.org/10.1016/j.compag.2022.107146

Knoll, F. J., Czymmek, V., Harders, L. O., & Hussmann, S. (2019). Real-time classification of weeds in organic carrot production using deep learning algorithms. *Computers and Electronics in Agriculture, 167*, 105097. https://doi.org/10.1016/j.compag.2019.105097

Le, V. N. T., Ahderom, S., & Alameh, K. (2020). Performances of the LBP based algorithm over CNN models for detecting crops and weeds with similar morphologies. *Sensors, 20*(8), 2193. https://doi.org/10.3390/s20082193

Milioto, A., Lottes, P., & Stachniss, C. (2017). Real-time blob-wise sugar beets vs weeds classification for monitoring fields using convolutional neural networks. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences,. https://doi.org/10.5194/isprs-annals-IV-2-W3-41-2017

Milioto, A., Lottes, P., & Stachniss, C. (2018). Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in CNNs. In *IEEE international conference on robotics and automation (ICRA)*, 2018 (pp. 2229–2235).

Moazzam, I. (2023). Tobacco and sesame crop datasets. Mendeley Datasets. https://data.mendeley.com/datasets/5dpc5gbgpz, https://data.mendeley.com/datasets/9pgv3ktk33

Moazzam, S. I., Khan, U. S., Qureshi, W. S., Nawaz, T., & Kunwar, F. (2023). Towards automated weed detection through two-stage semantic segmentation of tobacco and weed pixels in aerial Imagery. *Smart Agricultural Technology, 4*, 100142. https://doi.org/10.1016/j.atech.2022.100142

Nkemelu, D. K., Omeiza, D., & Lubalo, N. (2018). Deep convolutional neural network for plant seedlings classification. CoRR 1811.08404.

Partel, V., Kakarla, S. C., & Ampatzidis, Y. (2019). Development and evaluation of a low-cost and smart technology for precision weed management utilising artificial intelligence. *Computers and Electronics in Agriculture, 157*, 339–350. https://doi.org/10.1016/j.compag.2018.12.048

Sa, I., Popović, M., Khanna, R., Chen, Z., Lottes, P., Liebisch, F., Nieto, J., et al. (2018). WeedMap: A large-scale semantic weed mapping framework using aerial multispectral imaging and deep neural network for precision farming. *Remote Sensing, 10*(9), 1423. https://doi.org/10.3390/rs10091423

Sabzi, S., Abbaspour-Gilandeh, Y., & Arribas, J. I. (2020). An automatic visible-range video weed detection, segmentation and classification prototype in potato field. *Heliyon, 6*, 5. https://doi.org/10.1016/j.heliyon.2020.e03685

Sharpe, S. M., Schumann, A. W., & Boyd, N. S. (2020). Goosegrass detection in strawberry and tomato using a convolutional neural network. *Scientific Reports, 10*, 9548. https://doi.org/10.1038/s41598-020-66505-9

Subeesh, A., Bhole, S., Singh, K., Chandel, N. S., Rajwade, Y. A., Rao, K. V. R., Kumar, S. P., & Jat, D. (2022). Deep convolutional neural network models for weed detection in polyhouse grown bell peppers. *Artificial Intelligence in Agriculture, 6*, 47–54. https://doi.org/10.1016/j.aiia.2022.01.002

Wang, A., Peng, T., Cao, H., Xu, Y., Wei, X., & Cui, B. (2022). TIA-YOLOv5: An improved YOLOv5 network for real-time detection of crop and weed in the field. *Frontiers in Plant Science, 13*, 1091655. https://doi.org/10.3389/fpls.2022.1091655

Wendel, A., & Underwood, J. (2016). Self-supervised weed detection in vegetable crops using ground based hyperspectral imaging. In *IEEE international conference on robotics and automation (ICRA)*, Stockholm, Sweden, 2016 (pp. 5128–5135). https://doi.org/10.1109/ICRA.2016.7487717.

You, J., Liu, W., & Lee, J. (2020). A DNN-based semantic segmentation for detecting weed and crop. *Computers and Electronics in Agriculture, 178*, 105750. https://doi.org/10.1016/j.compag.2020.105750

Zhao, J., Tian, G., Qiu, C., Gu, B., Zheng, K., & Liu, Q. (2022). Weed detection in potato fields based on improved YOLOv4: Optimal speed and accuracy of weed detection in potato fields. *Electronics, 11*(22), 3709. https://doi.org/10.3390/electronics11223709