Towards Domain Generalization in Crop and Weed Segmentation for Precision Farming Robots

Jan Weyler Thomas Läbe Federico Magistri Jens Behley Cyrill Stachniss

***Abstract*—Precision farming robots offer the potential to reduce the amount of used agrochemicals through targeted interventions and thus are a promising step towards sustainable agriculture. A prerequisite for such systems is a robust plant classification system that can identify crops and weeds in various agricultural fields. Most vision-based systems train convolutional neural networks (CNNs) on a given dataset, i.e., the source domain, to perform semantic segmentation of images. However, deploying these models on unseen fields, i.e., in the target domain, often shows a low generalization capability. Enhancing the generalization capability of CNNs is critical to increasing their performance on target domains with different operational conditions. In this paper, we present a domain generalized semantic segmentation approach for robust crop and weed detection by effectively extending and diversifying the source domain to achieve high performance across different agricultural field conditions. We propose to leverage unlabeled images captured from various agricultural fields during training in a two-step framework. First, we suggest a method to automatically compute sparse annotations and use them to present the model more plant varieties and growth stages to enhance its generalization capability. Among others, we exploit unlabeled images from fields containing crops sown in rows. Second, we propose a style transfer method that renders the source domain images in the style of images from various fields to achieve increased diversification. We conduct extensive experiments and show that we achieve superior performance in crop-weed segmentation across various fields compared to state-of-the-art methods.**

***Index Terms*—Robotics and Automation in Agriculture and Forestry, Semantic Scene Understanding, Deep Learning for Visual Perception**

I. Introduction

An essential requirement for sustainable agriculture is to reduce the amount of agrochemicals used in farming, such as herbicides and pesticides, to decrease their negative impact on the environment [1], [2]. In this context, autonomous agricultural robots equipped with vision-based systems offer the potential to address this issue by deploying plant classification systems that automatically identify crops and weeds to perform targeted interventions in the field [3].

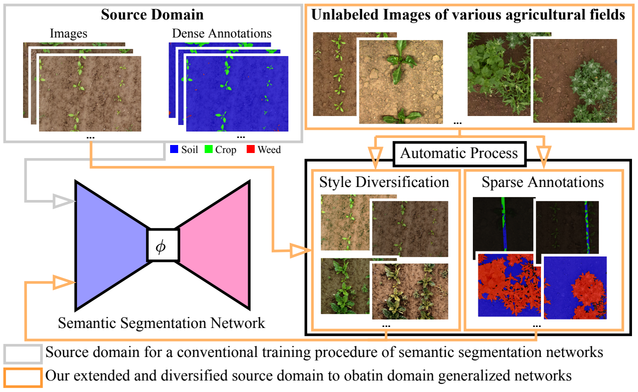
Source Domain
Images
Dense Annotations
...
...
🟦 Soil 🟩 Crop 🟥 Weed
Unlabeled Images of various agricultural fields
...
Automatic Process
Style Diversification
Sparse Annotations
...
...
ϕ
Semantic Segmentation Network
Source domain for a conventional training procedure of semantic segmentation networks
Our extended and diversified source domain to obatin domain generalized networks

Fig. 1: We propose a vision-based method to develop domain generalized semantic segmentation models by leveraging unlabeled images from various fields that reliably segments soil, crops, and weeds in arbitrary agricultural fields. We present methods to compute sparse annotation for these images automatically and to perform a style diversification of source domain images. We exploit both methods to train CNNs achieving high performance in various fields.

Most recent vision-based systems deploy convolutional neural networks (CNNs) to perform a semantic segmentation of soil, crops, and weeds in agricultural fields. Generally, these models are trained on a single dataset, i.e., the source domain, consisting of a set of images and its corresponding dense annotations, providing for each pixel its corresponding ground truth class. While achieving impressive results on images visually similar to the source domain, deploying these models on unseen agricultural fields, i.e., the target domain, results in a substantial performance decrease indicating a low generalization ability [4]. We attribute this to the domain gap between the source and target domain, i.e., the images of the source domain contain a limited view of plant varieties and growth stages and are restricted to a certain image style. Achieving domain generalization is essential for real-world deployment of precision farming robots operating in various fields [5]. In this paper, we address the issue of domain generalized semantic segmentation to reliably identify soil, crops, and weeds in images of arbitrary agricultural fields.

The main contribution of this paper is an end-to-end trainable pipeline for domain generalized semantic segmentation that achieves high performance in various agricultural fields. We leverage unlabeled images that we exploit during training to extend and diversify the source domain and increase the models' generalization capability, as sketched in Fig. 1. Our method does not need extra manual labeling and is simple to wrap into many network architectures. In sum, we make three key claims. First, we propose a method to leverage unlabeled images of various agricultural fields during training that achieves high generalization capabilities compared to conventionally trained semantic segmentation networks. Second, our framework performs superior to several domain adaptation

Manuscript received: Sep 9, 2022; Revised: Jan 11, 2023; Accepted: March 5, 2023. This paper was recommended for publication by Editor Hyungpil Moon upon evaluation of the Associate Editor and Reviewers' comments.

All authors are with the University of Bonn, Germany. Cyrill Stachniss is additionally with the Department of Engineering Science at the University of Oxford, UK, and with the Lamarr Institute for Machine Learning and Artificial Intelligence, Germany.

This work has partially been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy, EXC-2070 – 390732324 – PhenoRob and under STA 1051/5-1 within the FOR 5351 – AID4Crops.

methods that, contrary to ours, require image samples from the target domain. Lastly, our approach outperforms other domain generalization methods. Our implementation and datasets are available at: https://github.com/PRBonn/DG-CWS.

II. Related Work

Several vision-based approaches have been proposed for semantic segmentation in agricultural fields using handcrafted features [6], [7] or CNN-based methods [1], [2], [8]. Below, we aim to provide a broad overview of semantic segmentation methods and approaches to mitigate the performance decrease for unseen target domains.

Semantic Segmentation. Most current methods use CNNs for semantic segmentation of images from agricultural fields to predict a pixel-wise classification. Milioto et al. [9] propose a vision-based classification system based on CNNs to distinguish crops and weeds. Initially, they apply a preprocessing step to separate vegetation from soil and subsequently perform a classification on cropped regions to distinguish crops and weeds. McCool et al. [8] combine multiple lightweight CNN models to a mixture model to perform a fast and accurate weed segmentation on agricultural robots.

Contrary to our approach, these methods are trained on a single source domain and thus show low generalization capabilities when applied to various unseen target domains [4].

Domain Adaptation. Several works address the generalization issue by adapting the images of the source domain to the target domain [10], [4], [11]. Generally, these methods require image samples from the target domain to perform the adaptation. Subsequently, they train a CNN based on the adapted images of the source domain and deploy this model to the target domain to mitigate the decrease in performance.

Park et al. [11] generate images in the source domain that have the appearance of the target domain but explicitly preserve content by maximizing the mutual information between corresponding image patches with contrastive learning. Cherian et al. [10] emphasize that the semantic classes of the original and adapted image should be identical. Accordingly, they present a method that enforces semantic consistency during domain adaptation to achieve realistically adapted images. Gogoll et al. [4] propose an unsupervised domain adaptation for plant segmentation using generative adversarial networks. They assume a single dataset containing images and ground truth annotations as the source domain and image samples of the target domain without annotations. Next, they adapt images of the source domain to the target domain and train a CNN using the adapted dataset that achieves increased performance in the target domain.

These methods require images from the target domain to perform the adaptation, which limits their applicability.

Domain Generalization. In contrast, domain generalization [12] overcomes this limitation and aims at training robust CNNs that achieve high performance in arbitrary unseen domains. Contrary to domain adaption, any image of the target domain is unavailable before deploying the model.

Hendrycks et al. [13] propose a data augmentation procedure to diversify the source domain during training. Specifically, they compute multiple augmented versions of the original image and effectively combine them into a single augmented image. Their method achieves high generalization capability and performs robustly on multiple target domains. In contrast, Choi et al. [14] propose to enhance the generalization capability of CNNs by introducing an objective function to encourage the network to be invariant to the image style of the source domain. They exploit covariance matrices from feature maps and remove correlations related to variation in image style. Thus, their method is less sensitive to domain-specific styles but focuses on domain-invariant content.

Unlike our approach, these methods do either not exploit unlabeled images and eventually operate within the source domain or do not utilize automatically computed sparse annotations based on unlabeled images. We leverage images of various fields to increase domain generalization capabilities by using sparse annotations and exploiting various image styles.

III. Our Approach

The main objective of our approach is to develop a model based on CNNs that enables agricultural robots to perform a reliable semantic segmentation of the classes soil, crop, and weed in various agricultural fields, even when the source and target domain differ substantially. We propose a method to train domain generalized semantic segmentation models that achieve high performance on RGB images captured by unmanned aerial vehicles (UAVs) or unmanned ground vehicles (UGVs) across different agricultural fields.

First, we describe in Sec. III-A the conventional training procedure of semantic segmentation models and its deficiencies regarding domain generalization. We propose to address these issues by leveraging unlabeled images captured from various agricultural fields during training in a two-step framework. In Sec. III-B and Sec. III-C we present a method to automatically compute sparse annotations for these images and subsequently exploit them during training by computing objective functions on a subset of pixels, see Sec. III-D. This enables us to extend the intra-class content of the source domain to provide our model with wide plant varieties at different growth stages and soil conditions to improve its generalization capability. Furthermore, we argue that the image style of the source domain is severely restricted, thus facilitating overfitting to the source style. Consequently, we exploit the unlabeled images in Sec. III-E and propose a style transfer method based on a whitening and coloring transformation that diversifies the source domain by rendering its images in various real-world styles. In Sec. III-F, we exploit these images during training to alleviate overfitting. In sum, we utilize densely and sparsely annotated images and style-transferred versions of densely annotated images. Finally, we provide implementation details in Sec. III-G.

A. Supervised Semantic Segmentation

The key task of supervised semantic segmentation with CNNs is to train a network ϕ that provides for an RGB image $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ a corresponding prediction $\mathbf{p} \in \mathbb{R}^{H \times W \times K}$, which models pixel-wise a categorical distribution of K possible classes. Let H and W denote the height and

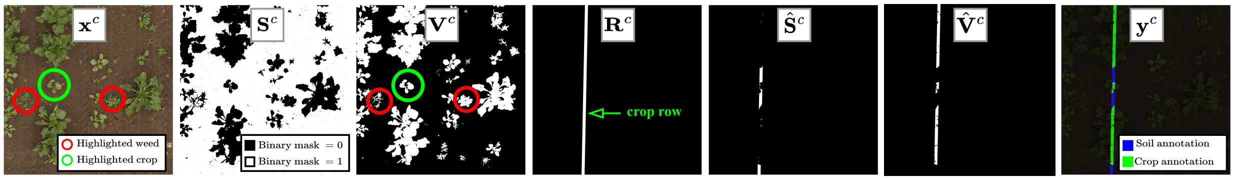

Fig. 2: Our method to automatically compute sparse annotations for soil and crops in images $\mathbf{x}^c$ of cultivated fields. First, we compute soil masks $\mathbf{S}^c$ and vegetation mask $\mathbf{V}^c$. However, these masks may still contain pixels associated to weeds. Thus, we detect a crop row, denoted as $\mathbf{R}^c$. Finally, we generate restricted masks $\hat{\mathbf{S}}^c$ and $\hat{\mathbf{V}}^c$ to consider only pixels along the row in the final sparse annotation $\mathbf{y}^c$.

width of an image, respectively. In supervised semantic segmentation, we have access to the source domain dataset $\mathcal{D}^s = \{(\mathbf{x}^s, \mathbf{y}^s)\}$ with $|\mathcal{D}^s| = N$, containing N pairs of RGB images $\mathbf{x}^s \in \mathbb{R}^{H \times W \times 3}$ and corresponding ground truth annotations $\mathbf{y}^s \in \mathbb{Z}^{H \times W \times K}$, which contain for each pixel a one-hot encoded vector over K classes. Thus, each image in the source domain is densely annotated. For our task, $K = 3$ since we define the classes soil, crop, and weed. During training, we follow best practice and minimize the cross-entropy objective to optimize the models' parameters:

$$\mathcal{L}^s_{\text{dense}}(\mathbf{y}^s, \mathbf{p}^s) = -\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{k=1}^{K}\mathbf{y}^s_{hwk}\log\left(\mathbf{p}^s_{hwk}\right), \quad (1)$$

where $\mathbf{p}^s = \phi(\mathbf{x}^s)$ represents the model prediction.

While deploying ϕ on images similar to the source domain results in impressive performance, it often fails to perform appropriately on other target domains due to the domain gap and its low generalization capability [14], [13]. We attribute this to a limited source domain with a restricted set of plant varieties, growth stages, and soil conditions, which occur in various target domains. Additionally, the model ϕ tends to overfit to the image styles provided in the source domain [14] since the set of $\mathbf{x}^s$ is captured with a limited variability of illumination. However, increasing the generalization capability by providing a large-scale source domain with dense annotations covering a wide range of plant varieties and image styles is not viable since the annotations are typically acquired manually and, thus, time- and labor-intensive.

Contrary, we propose an approach to leverage unlabeled, real-world images of various agricultural fields. Next, we present our method to automatically compute sparse annotations for these images that does not require manual intervention and subsequently exploit them during training to increase the generalization capabilities.

B. Sparse Annotations for Soil and Crops

First, we propose an approach to obtain sparse annotations for soil and crops in images of conventionally cultivated fields containing crops sown in rows by exploiting the spatial arrangement. In this work, we refer to sparse annotations as a subset of pixels with associated ground truth classes that we leverage during training to extend the source domain. First, we employ redness [15] and greenness [16] indices to obtain binary masks of pixels belonging to soil and vegetation. Subsequently, we detect a crop row by a Hough transform and exploit this information to filter out pixels belonging to weeds [17]. Finally, we obtain sparse annotations for pixels belonging to soil or crops, as shown in Fig. 2.

Similar to related work [15], [16], we first convert an RGB image $\mathbf{x}^c \in \mathbb{R}^{H \times W \times 3}$ of a conventionally cultivated crop field to the HSV color space and denote its hue channel as $\mathcal{H}^c$, where each pixel $\mathcal{H}^c_{hw} \in [0°, 360°)$. The value 0° indicates red pixels, 120° green pixels, and 240° blue pixels. Subsequently, we compute a binary soil mask $\mathbf{S}^c \in \mathbb{Z}^{H \times W}$ by a thresholding operation defined as:

$$\mathbf{S}^c_{hw} = \begin{cases} 1, & \text{if } \mathcal{H}^c_{hw} \leq 45° \vee \mathcal{H}^c_{hw} \geq 315° \\ 0, & \text{otherwise} \end{cases}, \quad (2)$$

and a binary vegetation mask $\mathbf{V}^c \in \mathbb{Z}^{H \times W}$:

$$\mathbf{V}^c_{hw} = \begin{cases} 1, & \text{if } 50° \leq \mathcal{H}^c_{hw} \leq 175° \\ 0, & \text{otherwise} \end{cases}. \quad (3)$$

Since our primary aim is to generate sparse annotations covering a fraction of all pixels, our method allows setting the thresholds conservatively to detect some pixels that are likely to belong to soil or vegetation but not all of them. Thus, we do not fine-tune the thresholds to stress that the method's heuristics are broadly applicable on crop row fields.

As highlight by red circles in Fig. 2 these masks may still contain undesired pixel belonging to weeds. Thus, we suggest to detect a single crop row in $\mathbf{V}^c$ by applying a Hough transform to constrain the vegetation pixels in $\mathbf{V}^c$ to crops and remove weeds. The Hough transform computes based on $\mathbf{V}^c$ for each parameter θ and r of a line in the polar system $l : y = -\tan(\theta)^{-1}x + r\sin(\theta)^{-1}$ its support in terms of vegetation pixels belonging to l, where $y \in [1, H]$ and $x \in [1, W]$. Note that we restrict the parameter space of θ to $\theta \leq 20° \vee \theta \geq 340°$ and thus consider only lines approximately vertical. This assumption is based on the fact that common farming robots capture images along vertical crop rows [18], [2]. Since the number of crop rows is variable, we suggest to identify the single most dominant line $\hat{l}$ with the most support in the Hough space. Therefore, we miss some rows, but we aim for correct annotations and do not want to introduce wrong labels.

Next, we compute a binary crop row mask $\mathbf{R}^c \in \mathbb{Z}^{H \times W}$:

$$\mathbf{R}^c_{hw} = \begin{cases} 1, & \text{if } d\left(\hat{l}, (h, w)\right) \leq \epsilon \\ 0, & \text{otherwise} \end{cases}, \quad (4)$$

where $d\left(\hat{l}, (h, w)\right)$ is the orthogonal distance between the line and a pixel location (h, w). Thus, $\mathbf{R}^c$ contains the crop row with a width of ϵ, see Fig. 2. Since our images $\mathbf{x}^c$ have a ground sampling distance (GSD) between $0.33\,\frac{\text{mm}}{\text{px}}$ and $1\,\frac{\text{mm}}{\text{px}}$, we set ϵ to $35\,\text{px}$ to cover a reasonable width between $11.5\,\text{mm}$ and $35\,\text{mm}$ for small and large crops. As before, we state that

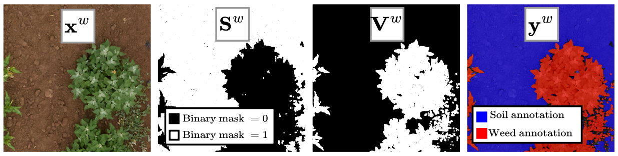

Fig. 3: Our method to automatically compute sparse annotations for soil and weed in images $\mathbf{x}^w$ of uncultivated fields. We exploit soil masks $\mathbf{S}^w$ and vegetation mask $\mathbf{V}^w$ to get sparse annotations $\mathbf{y}^w$.

a precise fine-tuning of ϵ is not required since we aim to obtain sparse annotations only.

Following, we constrain the binary masks $\mathbf{S}^c$ and $\mathbf{V}^c$ to pixels belonging to the crop row inherent in $\mathbf{R}^c$. Intuitively, since crops are cultivated along rows, we are confident that the restricted set of pixels belongs to soil or crops but not weeds. We define the constrained binary soil mask as $\hat{\mathbf{S}}^c$:

$$\hat{\mathbf{S}}^c_{hw} = \begin{cases} 1, & \text{if } \mathbf{R}^c_{hw} = 1 \land \mathbf{S}^c_{hw} = 1 \\ 0, & \text{otherwise} \end{cases} , \quad (5)$$

and similarly the constrained binary vegetation mask $\hat{\mathbf{V}}^c$:

$$\hat{\mathbf{V}}^c_{hw} = \begin{cases} 1, & \text{if } \mathbf{R}^c_{hw} = 1 \land \mathbf{V}^c_{hw} = 1 \\ 0, & \text{otherwise} \end{cases} . \quad (6)$$

We exploit both constrained binary masks to generate the sparse ground truth annotation $\mathbf{y}^c \in \mathbb{Z}^{H \times W \times K}$. We assign each pixel with $\hat{\mathbf{S}}^c_{hw} = 1$ to the class soil and each pixel with $\hat{\mathbf{V}}^c_{hw} = 1$ to the class crop, see Fig. 2. We define $\mathcal{A}^c$ as the set of automatically annotated pixels in $\mathbf{x}^c$ and emphasize that the annotations of the remaining pixels are undefined.

Ultimately, our proposed procedure enables us to generate an automatically, sparsely annotated soil-versus-crop dataset $\mathcal{D}^c = \{(\mathbf{x}^c, \mathbf{y}^c)\}$, with $|\mathcal{D}^c| = M$, based on M images captured from various conventionally cultivated fields that we exploit during training, see Sec. III-D. We tested our procedure on sugar beet fields, but we believe it is generic to any crop in cultivated fields sown in rows.

C. Sparse Annotations for Soil and Weeds

Additionally, we automatically compute sparse annotations for soil and weeds based on images of uncultivated fields, i.e., agricultural wastelands not containing any target crops.

Let $\mathbf{x}^w \in \mathbb{R}^{H \times W \times 3}$ be an RGB image of an uncultivated field. We compute the binary soil and vegetation mask as shown in Eq. (2) and Eq. (3). Consequently, we denote them as $\mathbf{S}^w$ and $\mathbf{V}^w$ and illustrate them in Fig. 3. However, in this case, we do not enforce any further constraint on the binary masks since $\mathbf{x}^w$ contains only soil and weed.

Next, we exploit the unconstrained binary masks to generate the sparse ground truth annotation $\mathbf{y}^w \in \mathbb{Z}^{H \times W \times K}$. We assign each pixel with $\mathbf{S}^w_{hw} = 1$ to the class soil and with $\mathbf{V}^w_{hw} = 1$ to the class weed, see Fig. 3. We define $\mathcal{A}^w$ as the set of automatically annotated pixels in $\mathbf{x}^w$ and stress that the annotations of the remaining pixels are undefined.

We compute these annotations for T images captured at various uncultivated fields to generate a sparsely annotated soil-versus-weed dataset $\mathcal{D}^w = \{(\mathbf{x}^w, \mathbf{y}^w)\}$ with $|\mathcal{D}^w| = T$. In the following, we exploit $\mathcal{D}^c$ and $\mathcal{D}^w$ during training.

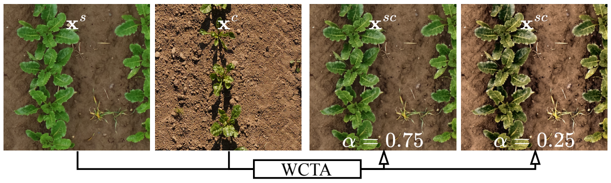

Fig. 4: Our proposed WCTA diversifies the style of source domain images $\mathbf{x}^s$, e.g., we transfer the style from an image $\mathbf{x}^c$ to $\mathbf{x}^s$, where α denotes the controllable transformation strength.

D. Training with Sparse Annotations

A key reason for the low generalization capability of conventionally trained CNNs is that the source domain contains limited source content [12], e.g., an insufficient amount of plant varieties and growth stages with constrained soil conditions. We address this issue by additionally involving the sparsely annotated datasets $\mathcal{D}^c$ and $\mathcal{D}^w$ during training.

Conventionally, a network ϕ processes an image $\mathbf{x}^s$ and computes pixel-wise an objective function to optimize the models' parameters by minimizing Eq. (1). Contrary, we additionally process an image $\mathbf{x}^c$ from $\mathcal{D}^c$ during training to obtain its predictions $\mathbf{p}^c = \phi(\mathbf{x}^c)$. Next, we propose to compute the cross-entropy objective effectively on the subset of pixels $\mathcal{A}^c$ belonging to sparse annotations:

$$\mathcal{L}^c_{\text{sparse}}(\mathbf{y}^c, \mathbf{p}^c) = -\frac{1}{|\mathcal{A}^c|} \sum_{i=1}^{|\mathcal{A}^c|} \mathbf{y}^c_i \log(\mathbf{p}^c_i) . \quad (7)$$

By minimizing Eq. (7), we present the network with extended content of real fields and enhance the generalization capabilities by enforcing the correct class for various crops in different soil conditions and prevent false weed predictions.

We further increase the intra-class content and process additionally an image $\mathbf{x}^w$ from $\mathcal{D}^w$. Like previously, we compute the cross-entropy function based on the set of sparsely annotated pixels in $\mathcal{A}^w$ and denote the corresponding objective as $\mathcal{L}^w_{\text{sparse}}(\mathbf{y}^w, \mathbf{p}^w)$. We minimize this objective to train our network on various weeds that occur in differing soil conditions and penalize false crop predictions.

E. Source Image Style Diversification

Another cause of low generalization capability is the often limited image style of the source domain [14] that does not cover the variety of illumination that occurs at various locations in images of real fields. Thus, we propose a method to diversify the style of images in the source domain. Specifically, we apply a whitening and color transformation (WCT) to transform an image in the source domain such that it inherits the covariance matrix [19] of an image in $\mathcal{D}^c$ or $\mathcal{D}^w$, which cover various real-world conditions. Chiu et al. [19] show that the WCT achieves good results among different approaches as it effectively transforms the style but preserves the content. Finally, we perform an alpha blending an denote our method as WCTA. Next, we describe the three-step procedure of our WCTA using an image pair $\mathbf{x}^s$ and $\mathbf{x}^c$.

Whitening Transformation. In this step, we employ a whitening transformation [19] to compute a representation of the image $\mathbf{x}^s$ such that its color channels are uncorrelated and have unit variance. First, we specify $\mathbf{f}^s \in \mathbb{R}^{3 \times HW}$

as a reshaped version of $\mathbf{x}^s$ that contains its RGB values row-wise. Before whitening, we center $\mathbf{f}^s$ by subtracting its mean $\mathbf{m}^s \in \mathbb{R}^{3\times 1}$ and denote its centered version as $\bar{\mathbf{f}}^s \in \mathbb{R}^{3\times HW}$. Let $\mathbf{Q}^s \in \mathbb{R}^{3\times 3}$ be an orthogonal matrix containing the eigenvectors of the covariance matrix $\mathbf{\Sigma}^s = \frac{1}{HW-1}\bar{\mathbf{f}}^s{\bar{\mathbf{f}}^s}^\top$ and $\mathbf{\Lambda}^s \in \mathbb{R}^{3\times 3}$ be the diagonal matrix of eigenvalues. We perform the whitening as following:

$$\hat{\mathbf{f}}^s = \mathbf{Q}^s {\mathbf{\Lambda}^s}^{-\frac{1}{2}} {\mathbf{Q}^s}^\top \bar{\mathbf{f}}^s, \tag{8}$$

such that $\hat{\mathbf{f}}^s{\hat{\mathbf{f}}^s}^\top = I_3$ holds true for $\hat{\mathbf{f}}^s \in \mathbb{R}^{3\times HW}$.

Coloring Transformation. Next, we aim to transfer the covariance matrix of a sparsely annotated image $\mathbf{x}^c$ to the result of the previous operation. This procedure provides a representation of the image $\mathbf{x}^s$ that is visually similar to $\mathbf{x}^c$ but preserves its content [19]. As previously, we denote $\mathbf{f}^c \in \mathbb{R}^{3\times HW}$ as the reshaped version of $\mathbf{x}^c$ and subtract its mean vector $\mathbf{m}^c \in \mathbb{R}^{3\times 1}$ to obtain a centered representation $\bar{\mathbf{f}}^c$. Let $\mathbf{Q}^c \in \mathbb{R}^{3\times 3}$ contain the eigenvectors of the covariance matrix $\mathbf{\Sigma}^c$ and $\mathbf{\Lambda}^c \in \mathbb{R}^{3\times 3}$ be the diagonal matrix of associated eigenvalues. Next, we perform the coloring:

$$\hat{\mathbf{f}}^{sc} = \mathbf{Q}^c {\mathbf{\Lambda}^c}^{\frac{1}{2}} {\mathbf{Q}^c}^\top \hat{\mathbf{f}}^s, \tag{9}$$

which essentially represents an inverse whitening. This transformation ensures that $\hat{\mathbf{f}}^{sc} \in \mathbb{R}^{3\times HW}$ has the desired correlations between its color channels, i.e., $\hat{\mathbf{f}}^{sc}{\hat{\mathbf{f}}^{sc}}^\top = \bar{\mathbf{f}}^c{\bar{\mathbf{f}}^c}^\top$. Thus, the transformed image $\hat{\mathbf{f}}^{sc}$ inherits the covariance matrix of the sparsely annotated image $\mathbf{f}^c$. Finally, we recenter the transformed image $\hat{\mathbf{f}}^{sc} = \hat{\mathbf{f}}^{sc} + \mathbf{m}^c$ and perform a reshaping to obtain the transformed image $\mathbf{x}^{sc} \in \mathbb{R}^{H\times W\times 3}$, see Fig. 4.

Alpha Blending. Next, we propose a method to combine an image $\mathbf{x}^s$ and its transformed representation $\mathbf{x}^{sc}$ by a parameter $\alpha \in [0, 1]$ that controlls the transformation strength:

$$\mathbf{x}^{sc} = \alpha \mathbf{x}^s + (1-\alpha)\,\mathbf{x}^{sc}. \tag{10}$$

Using $\alpha = 0$ exploits the color transformation entirely, $\alpha = 1$ preserves the original image $\mathbf{x}^s$, and intermediate values of α combine both representations proportionally. Since we aim to diversify the source domain, we randomly sample α from a uniform distribution in the interval $[0, 1]$, see Fig. 4. We perform the WCTA also for image pairs $\mathbf{x}^s$ and $\mathbf{x}^w$ since $\mathbf{x}^w$ covers yet another set of real-world conditions, see Fig. 5.

F. Training with Style-Diversified Source Images

Next, we exploit the WCTA during training to alleviate overfitting to the restricted source style. Particularly, we propose in Sec. III-D to process images $\mathbf{x}^c$ and $\mathbf{x}^w$ additionally to an image $\mathbf{x}^s$ to increase intra-class content variability based on their sparse annotations. Next, we additionally exploit $\mathbf{x}^c$ and $\mathbf{x}^w$ to compute style-diversified representations $\mathbf{x}^{sc}$ and $\mathbf{x}^{sw}$ of the source image that serve as additional inputs to the network, see Fig. 5. Consequently, we obtain their corresponding predictions $\mathbf{p}^{sc} = \phi\,(\mathbf{x}^{sc})$ and $\mathbf{p}^{sw} = \phi\,(\mathbf{x}^{sw})$ that are both $\in \mathbb{R}^{H\times W\times K}$. Since the style-diversified images $\mathbf{x}^{sc}$ and $\mathbf{x}^{sw}$ have the same semantic content as $\mathbf{x}^s$ (see Fig. 5) their ground truth annotation is $\mathbf{y}^s$. Thus, we compute the cross-entropy similiar as in Eq. (1) and denote the objective using the style-diversified image $\mathbf{x}^{sc}$ as $\mathcal{L}^{sc}_{\text{dense}}\,(\mathbf{y}^s, \mathbf{p}^{sc})$ and using $\mathbf{x}^{sw}$ as $\mathcal{L}^{sw}_{\text{dense}}\,(\mathbf{y}^s, \mathbf{p}^{sw})$. By minimizing these objectives, we enforce the network to capture correct semantic classes for images with the same content but different styles to achieve domain generalization.

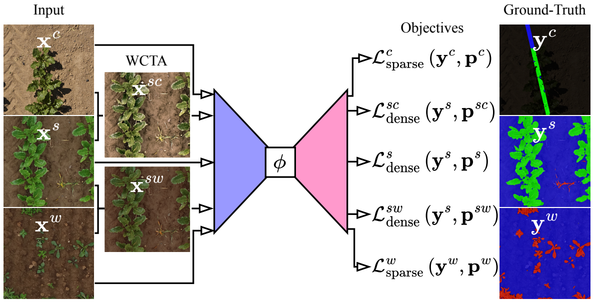
Input
$\mathbf{x}^c$
$\mathbf{x}^s$
$\mathbf{x}^w$
WCTA
$\mathbf{x}^{sc}$
$\mathbf{x}^{sw}$
ϕ
Objectives
$\mathcal{L}^c_{\text{sparse}}\,(\mathbf{y}^c, \mathbf{p}^c)$
$\mathcal{L}^{sc}_{\text{dense}}\,(\mathbf{y}^s, \mathbf{p}^{sc})$
$\mathcal{L}^s_{\text{dense}}\,(\mathbf{y}^s, \mathbf{p}^s)$
$\mathcal{L}^{sw}_{\text{dense}}\,(\mathbf{y}^s, \mathbf{p}^{sw})$
$\mathcal{L}^w_{\text{sparse}}\,(\mathbf{y}^w, \mathbf{p}^w)$
Ground-Truth
$\mathbf{y}^c$
$\mathbf{y}^s$
$\mathbf{y}^w$

Fig. 5: Our framework to train a semantic segmentation model ϕ that has high generalization capabilities. Besides training on images $\mathbf{x}^s$ and dense annotations $\mathbf{y}^s$ we suggest to process images $\mathbf{x}^c$ and $\mathbf{x}^w$ from various fields and propose an automatic procedure to compute their sparse annotations $\mathbf{y}^c$ and $\mathbf{y}^w$. Additionally, we perform a style transfer method (WCTA) to obtain images $\mathbf{x}^{sc}$ and $\mathbf{x}^{sw}$ with annotations $\mathbf{y}^s$ and exploit them also during training.

G. Implementation Details

We employ ERFNet [20] and DeepLabV3+ based on ResNet-50 [21], [22] as possible segmentation networks ϕ and report the results for both to show that our proposed method to increase the generalization capability is network-agnostic. The former aims at efficiency and has a small number of parameters, i.e., $2.1 \cdot 10^6$, while the latter is a high capacity architecture with $39.8 \cdot 10^6$ parameters. We train each model for 4096 epochs using the Adam optimization [23] with a batch size of 4 and a weight decay of $2 \cdot 10^{-4}$. At the initial 16 epochs, we linearly increase the learning rate to $1 \cdot 10^{-4}$ and subsequently apply a polynomial learning rate decay $\left(1 - \frac{e}{4096}\right)^3$, where e is the current epoch. During training, we randomly crop patches of size $768\,\text{px} \times 768\,\text{px}$ from each input image and use standard data augmentation methods, i.e., random adjustment of brightness, contrast, hue, and saturation, as well as random scaling and horizontal or vertical flipping. At each training step we pass a quintet of images, i.e., $\mathbf{x}_s$, $\mathbf{x}_c$, $\mathbf{x}_w$, $\mathbf{x}_{sc}$, and $\mathbf{x}_{sw}$, to our network and define the final loss function as a uniformly weighted sum of $\mathcal{L}^s_{\text{dense}}$, $\mathcal{L}^c_{\text{sparse}}$, $\mathcal{L}^w_{\text{sparse}}$, $\mathcal{L}^{sc}_{\text{dense}}$, and $\mathcal{L}^{sw}_{\text{dense}}$, as shown in Fig. 5. Thus, we set the weight of each loss equal to one to avoid overfitting to any specific domain.

During inference, we pass a single image to our network to obtain its associated semantic segmentation.

IV. Experimental Evaluation

In our experimental evaluation, we support our key claims that are: (i) We present a method that achieves high generalization capabilities compared to conventionally trained semantic segmentation networks, (ii) our framework performs superior to several domain adaptation methods, and (iii) our approach to exploit unlabeled images of various fields outperforms several domain generalization methods that eventually operate within the source domain.

TABLE I: Dataset Information.

	Bonn	$\mathcal{D}^c$	$\mathcal{D}^w$	Zurich	Stuttgart
# Images	379	2386	979	322	666
Platform	UAV	UAV/UGV	UAV/UGV	UAV	UGV
Annotations	dense	sparse	sparse	dense	dense
GSD $[\mathrm{mm/px}]$	1	0.33 - 1	0.33 - 1	1	0.33

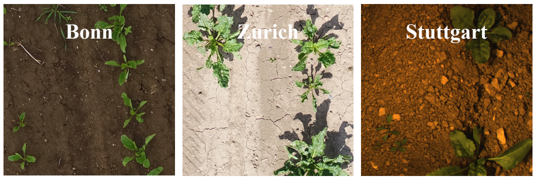

Fig. 6: Sample images from source und target domains.

Datasets. For the source domain $\mathcal{D}^s$, we deployed an UAV to collect a dataset from an agricultural field in Bonn that consists of RGB images captured over one month, see Tab. I. These images contain sugar beets between the 2- and 12 leaf growth stages and serve a substantial amount of different weed types. We use $70\,\%$ of these images for training, $15\,\%$ for validation, and define the remaining $15\,\%$ as test set to report the evaluation metrics. Additionally, we capture images from cultivated agricultural fields at various locations and deploy our proposed automatic procedure in Sec. III-B to compute sparse annotation for each image and generate the dataset $\mathcal{D}^c$ which is captured with different RGB cameras and cover a variety of sugar beets. Furthermore, we gather images from uncultivated fields at different locations containing various weeds with varying growth stages. We perform our procedure proposed in Sec. III-C to automatically compute sparse annotations for each image and generate the dataset $\mathcal{D}^w$. Finally, we employ $\mathcal{D}^s$, $\mathcal{D}^c$, and $\mathcal{D}^w$ to train a semantic segmentation network with our proposed method.

Additionally, we collected datasets from fields in Zurich and Stuttgart as target domains. Both contain sugar beets and weeds at various growth stages and differ visually substantially, see Fig. 6. As previously, we split each target domain into a training, validation, and testing split. Our method is not trained using any image in the target domains.

Evaluation Metric. To evaluate the performance of our approach and other methods, we compute the mean intersection over union (mIoU) $\in \mathbb{R}\,[0,1]$ across the classes soil, weed, and crop based on the test set of each dataset, where higher values indicate a better performance [24].

A. Comparison to Conventional Semantic Segmentation

To support our first key claim, we show that our approach achieves high generalization capabilities across different architectures compared to conventionally trained networks.

First, we use the densely annotated dataset of Bonn and perform a conventional training, i.e., we pass images of the training set to the network and optimize the model parameters by minimizing the objective in Eq. (1). Subsequently, we deploy the trained model on the test sets of Bonn, Zurich, and Stuttgart. We denote this approach as “conventional”. In Tab. II and Tab. III, we report the results based on DeepLabV3+ and ERFNet, respectively. In Tab. II, we see that the mIoU of the conventionally trained model is high with 0.92 on the test set of Bonn but decreases substantially on the test sets of Zurich and Stuttgart with 0.47 and 0.62, respectively. We attribute this to the domain gap between the source and target domain. In Tab. III, we observe similar behavior when using ERFNet, indicating that the issue of low generalization capability is rather network-agnostic.

TABLE II: Comparison of mIoU on the test sets of source and target domains between different approaches using DeepLabV3+.

Approach	Bonn (B) Source	Zurich (Z) Target	Stuttgart (S) Target	Avg.
Conventional [21]	**0.92**	0.47	0.62	0.67
Domain Adaption				
SCG [4] B → Z	0.86	0.69	0.36	0.64
SCG [4] B → S	0.49	0.36	0.59	0.48
CUT [11] B → Z	0.87	0.73	0.31	0.64
CUT [11] B → S	0.73	0.47	0.57	0.59
Domain Generalization				
AugMix [13]	0.91	0.77	0.67	0.78
RobustNet [14]	0.88	0.70	0.68	0.75
Ours	0.90	**0.79**	**0.78**	**0.82**
Upper Bound	0.92	0.88	0.92	0.91

TABLE III: Comparison of mIoU on the test sets of source and target domains between different approaches using ERFNet.

Approach	Bonn (B) Source	Zurich (Z) Target	Stuttgart (S) Target	Avg.
Conventional [20]	**0.93**	0.48	0.57	0.66
Domain Adaption				
SCG [4] B → Z	0.86	0.71	0.44	0.67
SCG [4] B → S	0.47	0.38	0.59	0.48
CUT [11] B → Z	0.86	0.76	0.54	0.72
CUT [11] B → S	0.75	0.48	0.57	0.60
Domain Generalization				
AugMix [13]	0.92	0.78	0.73	0.81
Ours	0.90	**0.79**	**0.78**	**0.82**
Upper Bound	0.93	0.86	0.92	0.90

Next, we train a model based on our proposed approach as described in Sec. III and deploy it to each test set. In Tab. II, we see that our method substantially increases the performance on the test sets of the target domains. Specifically, we achieve a mIoU of 0.79 for Zurich and 0.78 for Stuttgart. These results indicate a better generalization capability of our method as it achieves a consistent performance on images of different agricultural fields captured by UAVs or UGVs with different GSDs. Simultaneously, the performance drop on test images of the source domain is small compared to the conventional method (0.92 vs. 0.90). Thus, our increased generalization performance does not come at the expense of a major performance decrease for the source domain. As before, we observe the same effect in Tab. III when using ERFNet to emphasize that our method is network-agnostic.

Finally, we put the mIoU scores into a broader context and provide an upper bound for each dataset for comparison. To do that, we conventionally train three networks based on the training sets of Bonn, Zurich, and Stuttgart for each architecture. Then, we apply each model to its corresponding test set and report the performance, e.g., we deploy the model

TABLE IV: Effect of each step in our proposed framework to train with sparse annotations (An.), WCTA, or both in terms of mIoU on the test sets of source and target domains using DeepLabV3+.

Sparse An.	WCTA	Bonn	Zurich	Stuttgart	Avg.
		Source	Target	Target	
✓		0.89	0.58	0.58	0.68
	✓	**0.93**	0.61	0.72	0.75
✓	✓	0.90	**0.79**	**0.78**	**0.82**

trained on the training set of Zurich to the test set of Zurich. In this setting, the domain gap between source and target domain is marginal and thus, the results can be seen as upper bounds, as shown in Tab. II and Tab. III. Accordingly, the average upper bound across all three datasets based on DeepLabV3+ is 0.91, see Tab. II. While the conventional approach achieves 0.67, we obtain 0.82 and thus substantially improve the generalization capabilities of CNNs, see Fig. 7.

B. Comparison to Domain Adaptation

The next experiments show that our approach outperforms commonly used domain adaptation methods. We use SCG [4] and CUT [11] to train networks on images of Bonn adapted to Zurich or Stuttgart. We refer to the models trained on Bonn adapted to Zurich by B $\rightarrow$ Z or to Stuttgart by B $\rightarrow$ S.

First, we compare with models where the source domain is adapted to Zurich. Note that the conventional approach based on ERFNet performs poorly on Zurich, i.e., mIoU of 0.48, see Tab. III. Contrary, the models trained on the adapted source domain substantially improve the performance, i.e., mIoU of 0.76 and 0.71 for CUT and SCG. However, our domain generalized model outperforms both and achieves a mIoU of 0.79. Thus, our method outperforms the domain adaptation methods even in their targeted domain. Additionally, we observe for CUT and SCG a noticeable performance decrease on the original source domain, i.e., Bonn. Particularly, both methods achieve a mIoU of 0.86 compared to 0.93 for the conventional approach. This effect is less pronounced for our method, i.e., we achieve a mIoU of 0.90. Furthermore, the performance of CUT and SCG drop substantially when applied to Stuttgart, i.e., mIoU of 0.54 and 0.44. In sum, the domain adapted models perform only well on the domain they are adapted to. Contrary, we perform consistent across unseen agricultural fields, i.e., mIoU of 0.78 and 0.79 for Stuttgart and Zurich. The results in Tab. II support these conclusions when using DeepLabV3+.

Next, we compare with models trained on the source domain adapted to Stuttgart. While the conventional approach obtains in Tab. III a mIoU of 0.57 on Stuttgart, the performance increase for SCG and CUT is barely or not present, i.e., 0.59 and 0.57. We analyze the domain adapted images visually and find that their appearance is unrealistic. We attribute this to the difference in GSD between images of Bonn and Stuttgart. Similarly, Gogoll et al. [4] perform the domain adaptation only between images with the same GSD. This limitation results in underwhelming performance. Contrary, our method leverages images under various conditions and extends the source domain by images with various GSDs and achieves an increased mIoU of 0.78 for Stuttgart.

C. Comparison to Domain Generalization

In the last experiments, we show that our method achieves superior performance among different domain generalization methods to support our third key claim. Specifically, we compete with RobustNet [14] and AugMix [13]. Unlike our approach, these methods do not include real-world images outside the source domain during training but suggest procedures that eventually operate within the source domain. Note that the implementation of RobustNet is only available for DeepLabV3+ and thus, we only report its results in Tab. II.

In Tab. II, we show that our approach achieves an average mIoU of 0.82 across all datasets while RobustNet and AugMix obtain scores of and 0.75 and 0.78. Both baselines perform worse on Stuttgart with mIoU scores of 0.68 and 0.67 while our method achieves 0.78. We attribute this performance decrease to the different GSDs in the source domain of Bonn and target domain of Stuttgart. During training, the baseline models are restricted by the source domain containing plants with a GSD of $1\,\frac{\text{mm}}{\text{px}}$. Consequently, their performance suffers on images of Stuttgart with a substantially increased GSD of $0.33\,\frac{\text{mm}}{\text{px}}$. On the contrary, our approach effectively exploits additional images of various agricultural fields captured with different GSDs and thus shows increased performance. These results highlight that training on a single source domain is insufficient to achieve high domain generalization capabilities even with strong data augmentations like AugMix since it does not cover a sufficient intra-class variety. Our method overcomes this limitation by including automatically computed sparse annotations from diverse fields. We observe similar behavior in Tab. III, indicating that our method is network-agnostic.

D. Ablation Studies

A key contribution of our approach is to leverage unlabeled images from various agricultural fields in a two-step framework to increase the generalization capability of CNNs. To demonstrate the contribution of each step, we train and evaluate three networks using DeepLabV3+. We train the first model based on the objectives $\mathcal{L}^{s}_{\text{dense}}$, $\mathcal{L}^{c}_{\text{sparse}}$, and $\mathcal{L}^{w}_{\text{sparse}}$, i.e., it considers the sparse annotation but not the WCTA. Contrary, we train the second model based on $\mathcal{L}^{s}_{\text{dense}}$, $\mathcal{L}^{sc}_{\text{dense}}$, and $\mathcal{L}^{sw}_{\text{dense}}$, i.e., it considers the WCTA but not the sparse annotations. The third model considers all objectives and thus exploits our framework entirely. In Tab. IV, we report the evaluation metrics for each network and state that the first two models achieve an average mIoU of 0.68 and 0.75 across all datasets, respectively. Thus, both achieve superior performance compared to the conventional approach in Tab. II, i.e., 0.67. Consequently, each proposed method increases the generalization capability. However, the combination of both achieves the best average mIoU of 0.82, see Tab. IV.

Additionally, setting α in Eq. (10) to a constant value of zero slightly decreases the mIoU of the model only considering the WCTA from 0.75 to 0.74. If we only use $\mathcal{L}^{s}_{\text{dense}}$ together with $\mathcal{L}^{c}_{\text{sparse}}$ the average mIoU drops slightly to 0.67, whereas using only $\mathcal{L}^{w}_{\text{sparse}}$ together with $\mathcal{L}^{s}_{\text{dense}}$ results in a big performance decrease to 0.59 average mIoU.

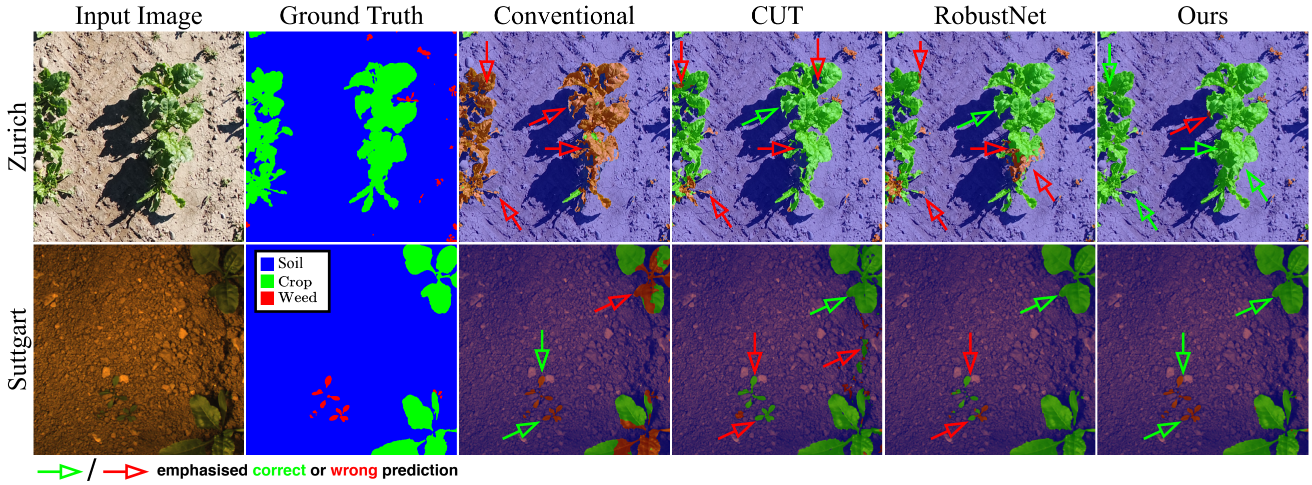

Fig. 7: Qualitative results based on DeepLabV3+ for the datasets Zurich and Stuttgart with different methods.

V. Conclusion

In this paper, we present a novel approach to leverage unlabeled images captured from various agricultural fields to develop domain generalized CNNs that enables agricultural robots to perform a reliable semantic segmentation of the classes soil, crop, and weed in different fields. First, we present a method to compute sparse annotations for these images automatically. Second, we propose a style transfer that renders images of the source domain in the style of real-world images captured in diverse conditions. We exploit both during training to increase the generalization capability of CNNs. We implemented and evaluated our approach based on multiple networks architectures and datasets. The experimental evaluation and comparisons with state-of-the-art methods support all claims made in this paper. We believe that our method allows to leverage vast amounts of unlabeled data to develop models with high generalization capabilities.

References

[1] P. Lottes, J. Behley, N. Chebrolu, A. Milioto, and C. Stachniss. Joint Stem Detection and Crop-Weed Classification for Plant-specific Treatment in Precision Farming. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2018.

[2] P. Lottes, J. Behley, N. Chebrolu, A. Milioto, and C. Stachniss. Robust Joint Stem Detection and Crop-Weed Classification using Image Sequences for Plant-Specific Treatment in Precision Farming. *Journal of Field Robotics (JFR)*, 37:20–34, 2020.

[3] M. Müter, P.S. Lammers, and L. Damerow. Development of an intra-row weeding system using electric servo drives and machine vision for plant detection. In *Proc. of the Agricultural Engineering Conf.*, 2013.

[4] D. Gogoll, P. Lottes, J. Weyler, N. Petrinic, and C. Stachniss. Unsupervised Domain Adaptation for Transferring Plant Classification Systems to New Field Environments, Crops, and Robots. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2020.

[5] D. Slaughter, D. Giles, and D. Downey. Autonomous robotic weed control systems: A review. *Computers and Electronics in Agriculture*, 61(1):63 – 78, 2008.

[6] S. Haug, A. Michaels, P. Biber, and J. Ostermann. Plant Classification System for Crop / Weed Discrimination without Segmentation. In *Proc. of the IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2014.

[7] P. Lottes, M. Höferlin, S. Sander, and C. Stachniss. Effective Vision-based Classification for Separating Sugar Beets and Weeds for Precision Farming. *Journal of Field Robotics (JFR)*, 34:1160–1178, 2017.

[8] C. McCool, T. Perez, and B. Upcroft. Mixtures of Lightweight Deep Convolutional Neural Networks: Applied to Agricultural Robotics. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2017.

[9] A. Milioto, P. Lottes, and C. Stachniss. Real-time Blob-wise Sugar Beets vs Weeds Classification for Monitoring Fields using Convolutional Neural Networks. In *Proc. of the Intl. Conf. on Unmanned Aerial Vehicles in Geomatics*, 2017.

[10] A. Cherian and A. Sullivan. Sem-GAN: semantically-consistent image-to-image translation. In *Proc. of the IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2019.

[11] T. Park, A. Efros, R. Zhang, and J. Zhu. Contrastive learning for unpaired image-to-image translation. In *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2020.

[12] S. Lee, H. Seong, S. Lee, and E. Kim. WildNet: Learning Domain Generalized Semantic Segmentation from the Wild. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[13] D. Hendrycks, N. Mu, E. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *Proc. of the Int. Conf. on Learning Representations (ICLR)*, 2019.

[14] S. Choi, S. Jung, H. Yun, J. Kim, S. .Kim, and J. Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[15] P. Owens and E. Rutledge. *Morphology. Encyclopedia of Soils in the Environment*. Elsevier, 2005.

[16] W. Yang, S. Wang, X. Zhao, J. Zhang, and J. Feng. Greenness identification based on HSV decision tree. *Information Processing in Agriculture*, 2(3-4):149–160, 2015.

[17] N. Chebrolu, T. Läbe, and C. Stachniss. Robust Long-Term Registration of UAV Images of Crop Fields for Precision Agriculture. *IEEE Robotics and Automation Letters*, 3(4):3097–3104, 2018.

[18] P. Lottes, J. Behley, A. Milioto, and C. Stachniss. Fully convolutional networks with sequential information for robust crop and weed detection in precision farming. *IEEE Robotics and Automation Letters (RA-L)*, 3:3097–3104, 2018.

[19] T. Chiu. Understanding generalized whitening and coloring transform for universal style transfer. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[20] E. Romera, J.M. Alvarez, L.M. Bergasa, and R. Arroyo. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Trans. on Intelligent Transportation Systems (ITS)*, 19(1):263–272, 2018.

[21] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2018.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[23] D. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *Proc. of the Int. Conf. on Learning Representations (ICLR)*, 2015.

[24] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez. A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv:1704.06857*, 2017.