IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

Unsupervised Generation of Labeled Training Images for Crop-Weed Segmentation in New Fields and on Different Robotic Platforms

Yue Linn Chong Jan Weyler Philipp Lottes Jens Behley Cyrill Stachniss

***Abstract*—Agricultural robots have the potential to improve the efficiency and sustainability of existing agricultural practices. Most autonomous agricultural robots rely on machine vision systems. Such systems, however, often perform worse in new fields or when the robotic platforms change. While we can alleviate the performance degradation by manually labeling more data obtained in the new setup, this procedure is labor and cost-intensive. Therefore, we propose an approach to improve the performance of machine vision systems for new fields and different robotic platforms without additional manual labeling. In an unsupervised manner, our approach can generate images and corresponding labels to train machine vision systems. We use StyleGAN2 to generate images that appear like they are from desired new field or robotic platform. Additionally, we propose a label refinement method to generate labels corresponding to the generated images. We show that our approach can improve the performance of the crop-weed segmentation task in new fields and on different robotic platforms without additional manual labeling.**

***Index Terms*—Robotics and Automation in Agriculture and Forestry, Deep Learning for Visual Perception, Object Detection, Segmentation and Categorization**

I. Introduction

AUTONOMOUS agricultural robots have the potential to improve the efficiency and sustainability of existing agricultural practices [9], [35]. However, autonomous agricultural robots' perception systems suffer performance degradation when domain shifts occur due to varying field conditions (e.g., weather, topsoil, or plant species) or changes to the robotic platform (e.g., sensor resolution, lighting, or sensor noise) [12].

Existing practices manually label images from the new domain to reduce performance degradation. However, constantly labeling new images is costly, time-consuming, and hard to scale. Thus, it can hinder real-world deployment. We aim to reduce labeling efforts by leveraging existing labeled images, referred to as the source images, to generate image-label pairs that are independent and identically distributed to the conditions the agricultural robots will operate in, referred to as the target domain, in an unsupervised manner. The generated image-label pairs replace the manually labeled images to train a new target domain network.

Manuscript received: Feb 24, 2023; Revised: May 25, 2023; Accepted: June 25, 2023. This paper was recommended for publication by Editor Hyungpil Moon upon evaluation of the Associate Editor and Reviewers' comments.

This work has partially been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy, EXC-2070 – 390732324 – PhenoRob. 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.

Digital Object Identifier (DOI): see top of this page.

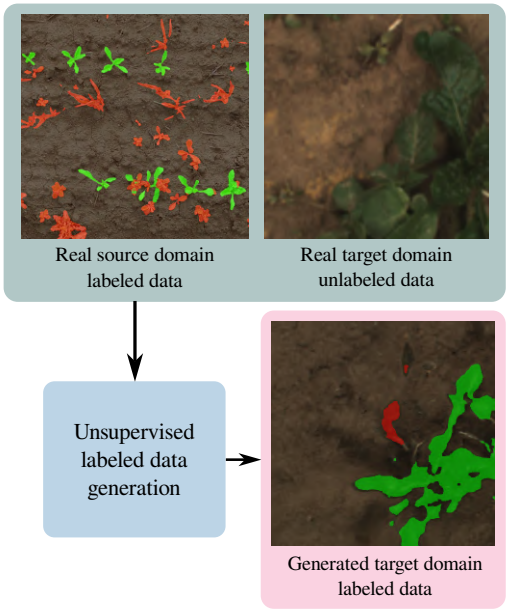


Fig. 1: We indicate image-labels pairs by overlaying the image with the labels, with crops as green and weeds as red. Using our unsupervised image-label pair generation approach, we are able to generate labeled images in the target domain by leveraging labeled source domain images and unlabeled target domain images.

One task affected by domain shifts is the crop-weed segmentation task, where we aim to distinguish between the crops, weeds, and non-vegetation for each pixel. Machine learning methods are used to perform crop-weed segmentation [23]. Several autonomous agricultural robots applications require the crop-weed segmentation such as automatic weeding [1], [40], phenotypic trait monitoring [25], [32], [39], and robot localisation [3]. Related segmentation tasks also occur in fruit picking [8]. We are interested in achieving this without requiring time-consuming manual labeling. Fig. 1 shows our overarching motivation for crop-weed segmentation.

Our unsupervised generation of image-label pairs in the target domain has two goals. First, we must generate images with the same quality and diversity as the target domain. Since the target domain can be arbitrary, our approach has to be agnostic to the field conditions or robots used in the source and the target domain. Our approach differs from existing methods [12] that assume similarity of the image resolution between the source and target domain. Second, we must ensure that the generated labels fit the corresponding images well. One can achieve a good fit by conditioning the image generation to fit a known label [12]. However, this is restrictive since the plant shapes may not be the same in the source and target domain. Therefore, we propose a weaker conditioning during image generation and refine the labels afterwards.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

The main contribution of this paper is a novel approach for generating crop-weed segmentation labels for images taken in agricultural fields, which fit the target domain in an unsupervised manner, i.e., without additional manual labeling. Our approach is applicable regardless of differences in image resolution, capture conditions, or field conditions between the source and target domain and improves the segmentation performance. While existing state-of-the-art methods focus mainly on the style transfer of the images, we also refine the generated labels to better match the generated images. We apply our approach and back up our claims with experiments using real-world agricultural data. Our data, code, and pre-trained weights are available at https://github.com/PRBonn/StyleGenForLabels.

II. Related Work

Existing work [2], [12], [22] has shown that unsupervised domain adaptation can improve the performance of several vision-based image interpretation tasks in a previously unseen target domain. In the paradigm of unsupervised domain adaptation, the target domain data is not labeled, so methods that perform data augmentation, e.g., the method proposed by Fawakherji et al. [10], is not applicable. Generally, unsupervised domain adaptation approaches aim to adapt the weights of a machine learning model trained on the source domain data to perform better on the target domain data without manually labeling the target domain data. While some works [36] propose novel mechanisms to adapt the weights, there are also frameworks [6], [12], [14] where they use data generated from variations of generative adversarial networks (GANs) to fine-tune the weights during training. For these works, the central idea is to use GANs to transfer data from the labeled source domain to the target domain so that the labels remain unchanged.

Rather than directly adapting the model weights to perform better in the target domain, our approach aims to generate image-label pairs that belong to the target domain. With the generated target domain image-label pairs, we can train for the downstream task network in a supervised manner. The advantage of this method is that it can train the downstream task network with any supervised training scheme, and there is no restriction on the architecture or size of the network. Moreover, we only need to generate the image-label pairs once, and we can reuse the same generated data if the target domain network architecture changes.

Several methods [6], [12], [14] use CycleGAN [43] to perform the unpaired style transfer between source and target data, and improve the downstream task performance in unlabeled target domains. Style transfer refers to transferring all visual aspects from the target image to the source image while maintaining the characteristics defined by the label of the source image. For example, in our use case, the style transfer would change the visual aspects such as lighting, soil texture, and colors while maintaining the original crop and weed positions in the generated target image. However, CycleGAN requires balancing four networks during training, which is difficult to achieve (each GAN has two networks: a generator network and a discriminator network). While there also exists methods that utilizes fewer networks, such as Label2Image-DA [21], which trains three networks simultaneously (i.e., two networks for the GAN and one downstream task network), theses approaches all share some limitations. Specifically, these methods use existing source domain labels to condition the generation of image-label pairs, which limits the diversity of the image-label pairs they can generate.

Our approach aims to perform unsupervised domain adaptation, which is agnostic to parameters such as the image resolution, or more pertinently, the ground sampling distance (GSD)[1] of the source and target domain.

Thus, our problem statement differs from that of super-resolution, such as the method proposed by Wang et al. [37] or Xu et al. [42], which only had image resolution or camera degradation as the domain gap. Our work aims to solve the similar problem to that of SRDA-Net [41], which adapts different domains of aerial images of different resolutions. However, in SRDA-Net, the difference in visual appearance between the domains is less drastic than in our use case, featuring different lighting and field conditions.

Beyond the paradigm of unsupervised domain adaptation, GANs can perform both, style transfer and resolution adaptation respectively. Specifically, methods built on StyleGAN [17] and StyleGAN2 [18] can perform style transfer via style mixing and super-resolution [5], [7], [15], [28], [34]. However, methods such as StyleRig [34] and StyleFusion [15] require known semantics of the target image, either by labels or a pre-trained auxiliary network. However, these semantics are unknown in unsupervised domain adaptation, and unsupervised domain adaptation aims to obtain these semantics for the target domain. Thus, it is not possible to adapt these methods directly. We also do not have access to additional information such as 3D data that SofGAN [5] uses to train a semantic segmentation model.

Our approach follows the idea of performing style transfer from images in the source domain to images in the target domain while maintaining the condition of the original source label [6], [12], [14]. Unlike existing work, we focus on the specific use case where the source and target domain have different field conditions and GSDs. In contrast to existing methods such as Label2Image-DA [21] and other CycleGAN-based methods [6], [12], [14], our method trains the GAN networks and the downstream task network separately. We also exploit StyeGAN2's style mixing to perform the style transfer to handle the difference in GSD between the source and target domain.

III. Our Approach

Fig. 2 shows the workflow of our proposed approach. It aims to generate pairs of images and generated labels belonging to the target domain, which we can use to supervise the training of a machine learning model and improve the model's

[1]We refer to GSD as the distance between the center of each pixel when projected onto the ground plane, which is calculated as the ratio of the height of ground plane to the distance of the center with respect to the projection center of the camera sensor [11].

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

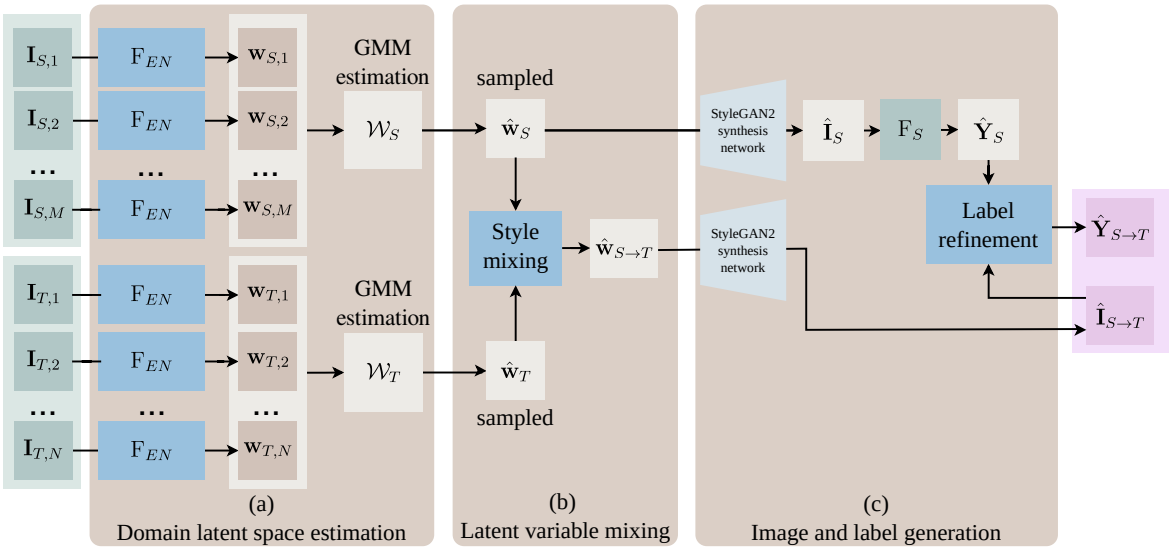


Fig. 2: Overall workflow of our approach. We indicate the inputs in green and outputs of our approach in magenta. Firstly, we estimate the distribution of the latent space $\mathcal{W}$ representing the source domain $\mathcal{W}_S$ and target domain $\mathcal{W}_T$. Secondly, we sample latent variables $\hat{\mathbf{w}}_S \sim \mathcal{W}_S$ and $\hat{\mathbf{w}}_T \sim \mathcal{W}_T$ and perform style mixing to obtain $\hat{\mathbf{w}}_{S\to T}$. Finally, we generate the output image, $\hat{\mathbf{I}}_{S\to T}$, and its labels, $\hat{\mathbf{Y}}_{S\to T}$.

performance in the target domain. To generate an image, we must first obtain the mixed latent variable corresponding to the generated image using style mixing. To generate the label, we perform label refinement on the label of the corresponding source image.

In the following, we denote variables belonging to the source domain with a subscript "S" and variables in the target domain, i.e., the domain we want to operate in, with a subscript "T". We denote an image from the source domain dataset with M images as $\mathbf{I}_{S,i}$ and its label as $\mathbf{Y}_{S,i}$, where $i \in \{1, 2, \ldots, M\}$. We denote images from the target dataset containing N images as $\mathbf{I}_{T,j}$, where $j \in \{1, 2, \ldots, N\}$.

Our approach has three parts: (a) source and target domain latent space estimation, (b) latent variable mixing, and (c) paired image and label generation, as illustrated in Fig. 2.

In the first part (a), we estimate two distributions in the latent space, which represent the source domain $\mathcal{W}_S$ and target domain $\mathcal{W}_T$. In the second part (b), we sample latent variables $\hat{\mathbf{w}}_S \sim \mathcal{W}_S$ and $\hat{\mathbf{w}}_T \sim \mathcal{W}_T$ and perform style mixing to obtain $\hat{\mathbf{w}}_{S\to T}$. In the final part (c), we generate the output image, $\hat{\mathbf{I}}_{S\to T}$, from $\hat{\mathbf{w}}_{S\to T}$ and its corresponding generated label, $\hat{\mathbf{Y}}_{S\to T}$ with the label refinement process.

A. Source and Target Domain Latent Space Estimation

The goal of the first part of our approach is to estimate the distribution in the latent space $\mathcal{W}$ of the source domain, $\mathcal{W}_S$, and that of the target domain, $\mathcal{W}_T$, which are in the latent space of StyleGAN2 [18]. StyleGAN2 is a generative adversarial network [13], which aims to generate images from the same distribution as the training images. As shown in Fig. 3, StyleGAN2 consists of two networks: the mapping network and the synthesis network. The mapping network maps a variable $\mathbf{z} \sim \mathcal{N}(0, I)$ to a latent variable $\mathbf{w} \in \mathcal{W}$. The synthesis network uses the latent variable $\mathbf{w}$ to generate an RGB image. We train StyleGAN2 to generate images from the source and target domains and freeze the weights of the mapping and the synthesis networks. Thus, we map $\mathcal{W}$ to the source and target domain images via the synthesis network.

We choose StyleGAN2 because training is relatively stable and it does not suffer mode collapse. StyleGAN2 is also able to generate diverse images of plants at different GSDs, growth stages, soil textures, and environmental conditions while generating photorealistic images with visually convincing textures. Moreover, we would like to take advantage of the disentangled latent space of StyleGAN2, which enables intuitive latent variable mixing, so-called style mixing [17].

To estimate $\mathcal{W}_S$ and $\mathcal{W}_T$, we first train an encoder network, F_{EN}, to perform style inversion, i.e., to map a given RGB image to the latent space as shown in Fig. 3. Similar to StyleRig [34], we train F_{EN} using the StyleGAN2 outputs. We generate the desired output, $\mathbf{w}$, of F_{EN}, using the mapping network from the sampled input $\mathbf{z}$. The input to F_{EN} is an image generated with the synthesis network from $\mathbf{w}$. We train F_{EN} to minimize the L2 distance between the predicted latent variable $\hat{\mathbf{w}}$ and desired output $\mathbf{w}$. In our implementation, we use pSp [28] for the encoder F_{EN}. Note that, unlike in pSp, our training method allows us to train the encoder without input-output image pairs.

With F_{EN}, we map each image $\mathbf{I}_{S,i}$ to a corresponding latent variable, $\mathbf{w}_{S,i}$, for $i = \{1, 2, \ldots, M\}$. We estimate $\mathcal{W}_S$ by parameterizing $\mathcal{W}_S$ as a Gaussian mixture model (GMM) with k components. We repeat the process for the target domain where we map $\mathbf{I}_{T,j}$ to the latent variable $\mathbf{w}_{T,j}$ for $j = \{1, 2, \ldots, N\}$ and estimate $\mathcal{W}_T$ as a GMM with k components.

B. Latent Variable Mixing

The objective of the latent variable mixing is to obtain the latent variable $\hat{\mathbf{w}}_{S\to T}$, which corresponds to an image $\hat{\mathbf{I}}_{S\to T}$. There are two conditions we want $\hat{\mathbf{I}}_{S\to T}$ to follow. First, we want $\hat{\mathbf{I}}_{S\to T}$ to be from the target domain. Second, $\hat{\mathbf{I}}_{S\to T}$ should have crops and weeds in the same locations as in $\hat{\mathbf{I}}_S$, such that $\hat{\mathbf{Y}}_S$ is similar to $\hat{\mathbf{Y}}_{S\to T}$. Ideally, we would like $\hat{\mathbf{Y}}_{S\to T}$ to

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

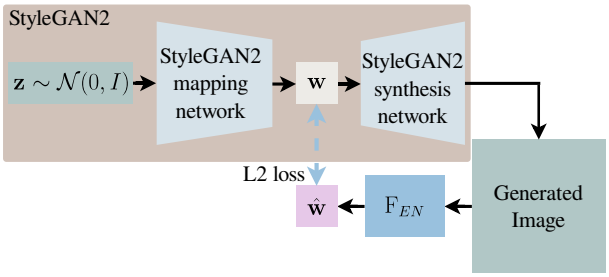


Fig. 3: We train the encoder, F_{EN}, to perform style inversion by first sampling a variable $\mathbf{z}$ as input for the mapping network, which maps $\mathbf{z}$ to latent space $\mathcal{W}$. With the latent variable $\mathbf{w}$, we generate the corresponding image with the synthesis network and input the generated image to F_{EN}. We use the original $\mathbf{w}$ as the ground truth of the output of F_{EN} and train F_{EN} with an L2 loss.

be equal to $\hat{\mathbf{Y}}_S$ so that we can transform $\hat{\mathbf{Y}}_S$ into $\hat{\mathbf{Y}}_{S\to T}$, in the label refinement process. We explain the label refinement process in the following Sec. III-C. We can obtain $\hat{\mathbf{w}}_{S\to T}$ that encodes $\hat{\mathbf{I}}_{S\to T}$ that follows this two conditions by style mixing latent variables $\hat{\mathbf{w}}_S$ and $\hat{\mathbf{w}}_T$, which encodes all the information required to generate $\hat{\mathbf{I}}_S$ and $\hat{\mathbf{I}}_T$.

In StyleGAN2, each latent variable $\mathbf{w}$ comprises multiple components and each component encodes different aspects of the generated image. In our implementation of StyleGAN2, each latent variable has 16 components, where $\hat{\mathbf{w}}_S = [{}^0\hat{\mathbf{w}}_S, {}^1\hat{\mathbf{w}}_S, \ldots, {}^{15}\hat{\mathbf{w}}_S]$ and similarly $\hat{\mathbf{w}}_T = [{}^0\hat{\mathbf{w}}_T, {}^1\hat{\mathbf{w}}_T, \ldots, {}^{15}\hat{\mathbf{w}}_T]$. Based on the architecture of StyleGAN2, we categorize the features into coarse-, medium-, and fine-grained feature components. The coarse- and medium-grained feature components encode the content of the generated image, such as the position and type of objects present. The fine-grained feature components encode fine features of the generated image, such as the overall visual aesthetic of the image, e.g., lighting condition or colors. In our approach, the fine-grained components are the 7^{th} to the 15^{th} components of the latent variable. To perform the style mixing [17], we take the components corresponding to the coarse- and medium-grained features from the latent variable representing the desired image content and that of the fine-grained features from the desired image visual aesthetic. Specifically, we swap the 7^{th} to the 15^{th} components of $\hat{\mathbf{w}}_S$ with that of $\hat{\mathbf{w}}_T$ to obtain $\hat{\mathbf{w}}_{S\to T} = [{}^0\hat{\mathbf{w}}_S, {}^1\hat{\mathbf{w}}_S, \ldots, {}^5\hat{\mathbf{w}}_S, {}^7\hat{\mathbf{w}}_T, {}^8\hat{\mathbf{w}}_T, \ldots, {}^{15}\hat{\mathbf{w}}_T]$.

With this, our approach can condition the generation of images by StyleGAN2, even though we do not have source-target paired data.

C. Paired Image and Label Generation

In the final part, we aim to finally generate the output image, $\hat{\mathbf{I}}_{S\to T}$, and generated label, $\hat{\mathbf{Y}}_{S\to T}$, using the outputs from the previous parts. To generate $\hat{\mathbf{I}}_{S\to T}$, we use the synthesis network to map $\hat{\mathbf{w}}_{S\to T}$ to the RGB image space.

To generate $\hat{\mathbf{Y}}_{S\to T}$, we perform the following steps. First, we train the supervised downstream task network F_S using images $\mathbf{I}_S$ and labels $\mathbf{Y}_S$. Second, we obtain $\hat{\mathbf{I}}_S$ from the synthesis network using $\hat{\mathbf{w}}_S$. Third, we obtain the predicted label $\hat{\mathbf{Y}}_S$ using F_S on $\hat{\mathbf{I}}_S$. Since we condition the latent variable mixing such that $\hat{\mathbf{I}}_{S\to T}$ has the same content as $\hat{\mathbf{I}}_S$,

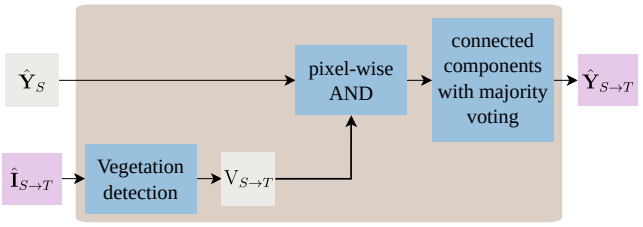


Fig. 4: We refine the label $\hat{\mathbf{Y}}_S$ to obtain the label $\hat{\mathbf{Y}}_{S\to T}$. We correct the plant boundaries in $\hat{\mathbf{Y}}_S$ using the vegetation mask of $\hat{\mathbf{I}}_{S\to T}$.

$\hat{\mathbf{Y}}_S$ should be equal to $\hat{\mathbf{Y}}_{S\to T}$. However, we found that $\hat{\mathbf{I}}_{S\to T}$ may not fit $\hat{\mathbf{Y}}_S$ well, particularly near the plant boundaries. The discrepancy may be due to the differences in GSD and, consequently, the plant sizes in the source and target domains. Additionally, we also noticed that the shape of the leaves may also vary between domains which further worsens the fit of $\hat{\mathbf{Y}}_S$ for $\hat{\mathbf{I}}_{S\to T}$. Therefore, we perform an additional step of label refinement to obtain $\bar{\mathbf{Y}}_{S\to T}$, which improves $\hat{\mathbf{Y}}_S$ to better correspond to $\hat{\mathbf{I}}_{S\to T}$.

The goal of the label refinement, shown in Fig. 4, is to correct the boundaries of the predicted label, $\bar{\mathbf{Y}}_S$, to better fit the plants in $\hat{\mathbf{I}}_{S\to T}$. To this end, we, first, separate the foreground (i.e., crops and weeds) and background (i.e., non-vegetation) in $\hat{\mathbf{I}}_{S\to T}$. The separation can be done using a Convolutional Neural Network vegetation classifier [24] or heuristics. In our approach, we perform the separation heuristically using a hue filter by thresholding out green pixels as vegetation followed by a morphological closing. We label pixels in $\hat{\mathbf{Y}}_{S\to T}$ as crops only if the same pixels are also crops in $\hat{\mathbf{Y}}_S$ and also marked as vegetation with the hue thresholding of $\hat{\mathbf{I}}_{S\to T}$. Then, we perform a connected component analysis on $\hat{\mathbf{Y}}_{S\to T}$. For components with multiple labels, we use majority voting to obtain the final $\hat{\mathbf{Y}}_{S\to T}$.

IV. Experimental Evaluation

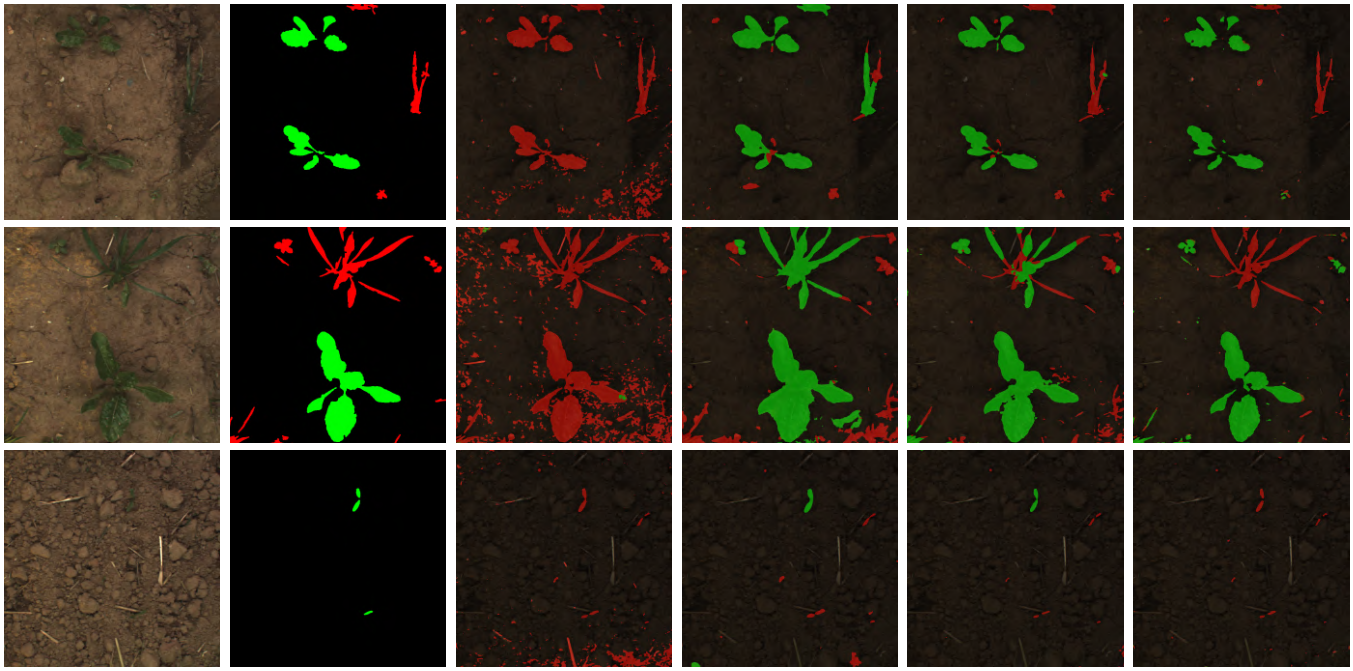
The main focus of this work is an unsupervised method to generate crop-weed segmentation labels for images in the target domain using labeled source domain images and unlabeled target domain images. The source domain can have different image GSDs, capture conditions, and field conditions than the target domain. The experiments support our claim that our approach improves the performance of crop-weed segmentation methods on the target domain despite differences in the source and target domain.

A. Experimental Setup and Parameters

For all experiments, we used a StyleGAN2 implemented in PyTorch [27] pre-trained on over 117k unmanned ground vehicle (UGV) images from the SugarBeets dataset [4]. The StyleGAN2 was then fine-tuned on the source and target images for 460k steps with batch size 32. We resize all images to 512×512 pixel.

As the source domain, we use low-resolution aerial vehicle (UAV) images with a GSD of 1 $\frac{\text{mm}}{\text{pixel}}$. The dataset has 379 labeled images, of which we use 255 images for training, 61 for validation, and 63 for testing. We reduce the difference in the source and target labels by upsampling 3× the source

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.



(a) Input image (b) Ground truth (c) Source (d) CycleGAN [43] (e) Gogoll et al. [12] (f) Ours

Fig. 5: Qualitative results of crop-weed segmentation. We indicate image-label pairs by overlaying the image with crops for green and weeds for red. We darken the pixels segmented as non-vegetation. We show the qualitative results for the semantic segmentation network trained on the source image-label pairs only in (c). In columns (d) and (e), we show the results when the semantic segmentation network trained on image-label pairs using CycleGAN [43] and Gogoll et al. [12]. Finally, we show the results of our method in column (f).

images and labels using Lanczos interpolation [20]. The up-sampled source images do not need to have the same GSD as the target images, as we perform the label refinement step. The synthesis network can also generate different GSDs for the source and target domains.

The target domain consists of UGV images collected at a different time of the year. The GSD of the images ranged from 0.3 $\frac{\text{mm}}{\text{pixel}}$ to 0.8 $\frac{\text{mm}}{\text{pixel}}$. In total, there are 926 images, of which we use 140 for testing for all scenarios, 647 for training, and 139 for validation. We used $k = 100$ components for the GMM to model the latent spaces. For the label refinement, we use a closing kernel of 5 pixels, and we threshold the hue in the range 40 to 140. We sample $Q = 15$k latents from $\mathcal{W}_S$ and $\mathcal{W}_T$, respectively.

B. Crop-Weed Segmentation Performance

The first experiment evaluates the performance of our approach on the downstream task of crop-weed segmentation to differentiate sugar beet crops from weeds. The experiment supports our claim that our approach improves the performance of a crop-weed segmentation network in the target domain, despite the differences in GSD, image capture, and field conditions of the source and target domains.

Specifically, we use images captured using an aerial vehicle, i.e., UAV as the source domain, and images collected using a ground vehicle, i.e., UGV as the target domain (see Sec. IV-A). We chose these domains because their images have different GSDs due to the difference in the cameras used and the distance of the camera from the crops. Furthermore, the UGV has artificial lighting, resulting in darker images than UAV images with natural lighting.

To perform the downstream task, we adopt the semantic segmentation network ERFNet [30] as F_S. ERFNet is an efficient and accurate architecture frequently used and achieves competitive performance in the agricultural domain [29], [38]. We repeat the training of ERFNet three times and report the average performance. For our approach, we train ERFNet from scratch since fine-tuning from the weights trained on the source domain does not improve the network's performance.

To show the performance without a domain gap, we report the performance of ERFNet when trained and tested on UAV captured images. We also report the performance of ERFNet trained and tested on UGV captured images showing the performance in the so-called "oracle case" where manual labels are available. These two cases' performance is the upper bound since manual labels are available. To show the performance degradation caused by the domain gap between the UAV and UGV, we report the performance of ERFNet trained on UAV captured images but tested on UGV captured images. For comparison, we also report the performance of Gogoll et al. [12] and CycleGAN [43]. For both these approaches, we tuned the learning rate to $5 \cdot 10^{-6}$ with a linearly decreasing scheduler after the first 100 epochs. We weigh the classes with 1, 2, and 4 for soil, crop, and weed, respectively, to account for the unbalanced number of pixels representing each class. We ran the training for 400 epochs. Finally, we report the performance of our method, where we train ERFNet on generated images and generated labels.

Tab. I shows the results of the crop-weed segmentation task. We use the mean intersection over union (mIoU) to show the performance of each approach. We also present the corresponding qualitative results in Fig. 5.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

TABLE I: Performance of crop-weed segmentation. The UAV domain is the source domain, and the UGV domain is the target domain. Our approach has best performance with the highest average mIoU for the domain adaptation cases.

Training dataset	Test dataset	mIoU / %			
		non-vegetation	crop	weed	average
UAV	UAV	98.6	91.9	70.2	86.9
UGV	UGV	99.5	86.4	65.5	83.8
UAV	UGV	89.2	14.6	8.99	37.6
CycleGAN [43]	UGV	96.0	47.7	12.2	52.0
Gogoll et al. [12]	UGV	98.2	55.6	**30.1**	61.3
Ours	UGV	**99.3**	**61.7**	25.9	**62.3**

TABLE II: Performance of StyleGAN2 [18] and baseline in generating in-domain images. Higher values mean better performance. Improved precision and density indicate the fidelity of generated images. Improved recall and coverage indicate the diversity of generated images. The UGV domain is the target domain.

Method	Improved precision	Improved recall	Density	Coverage
Gogoll et al. [12]	**0.0711**	0.158	0.0158	0.0227
Ours	0.0609	**0.271**	**0.0210**	**0.125**

Our results show that ERFNet performs well when the training and test images are both from UAV captured images or UGV captured images, with an average mIoU of 86.9 % and 83.8 %, respectively. There is a drop in mIoU when the training and test images have a domain gap, as shown by our results where we train ERFNet on UAV captured images and test on UGV captured images compared to when we test on UAV captured images, from the 86.9 % to 37.6 %.

Compared to where we only train using UAV images, our proposed method improves the performance of the crop-weed segmentation task on UGV images. We also show that we perform better than the method of Gogoll et al. [12] and CycleGAN [43] in the average mIoU and the non-vegetation and crop classes. As visible in the qualitative results in Fig. 5, our approach performs better at the segmentation of crops, especially at the plant borders. However, the method by Gogoll et al. [12] outperforms our method for the weed class. The performance of the weed class is more sensitive to errors since there are fewer weed instances in the dataset than in the other classes. The poor performance in the weed class of our proposed method is probably due to the misclassification of younger growth stage crops as weeds, as shown in Fig. 5. One possible reason why the ERFNet trained with our method performs worse when the crops are at an earlier growth stage is that we train StyleGAN2 on fewer images of earlier growth stage and is less likely to generate images with earlier growth stage crops.

In summary, our approach improves the performance of ERFNet without additional manual labels.

C. Generative Performance

The second experiment evaluates the generative performance and illustrates that our approach can generate labels that match the generated images from the target domain. We hypothesize that the better the generated images are representative of the UGV domain, the better the performance of the downstream tasks trained with the generated images.

(a) UGV images (b) Gogoll et al. [12] (c) Ours

Fig. 6: Qualitative results of generated images. Our generated images are able to replicate the size and color of the crops better than that of Gogoll et al.

We compare our approach to the CycleGAN-based method by Gogoll et al. [12] for generating images from the UGV domain. From the three runs used in Sec. IV-B, we report the evaluation on the images that yielded the best mIoU.

We measure how representative our images are to the UGV domain using metrics commonly used to evaluate generated image quality in methods such as GANs. Specifically, we evaluated our approach using the StudioGAN's [16] implementation of improved precision and recall [19], and density and coverage [26] metrics. The newer metrics of density and coverage are more representative of the performance of generating images compared to the older methods [26], so we will discuss the performance of our method based on these metrics. However, we also report the improved precision and improved recall of all approaches for completeness.

To briefly summarize the intuition behind the metrics, improved precision and recall are, respectively, a pair of values that represent different aspects. Density and improved precision represent the quality or fidelity of the generated images, while coverage and improved recall represent the diversity or variation of the images generated. We evaluate these metrics using InceptionV3 [33] weights trained on ImageNet [31], which may not be ideal in our agricultural use case. Tab. II shows the performance of our approach, and that of Gogoll et al. [12]. Fig. 6 shows the generated images alongside the UAV images for qualitative comparison. While our approach and Gogoll's approach are able to replicate the soil of the UGV images, Gogoll's approach is not able to correctly generate plants of the correct size or color. In contrast, our approach is able to generate plants of similar size and color to that of the UGV images.

Our approach shows superior density and coverage metrics performance compared to Gogoll et al. [12]. While it is possible to increase the improved precision score of StyleGAN2 methods using the "truncation trick" [18], we did not perform any truncation to avoid corrupting the style transfer.

The qualitative results show that the approach by Gogoll et al. [12] has difficulties in the coloration or texture of the

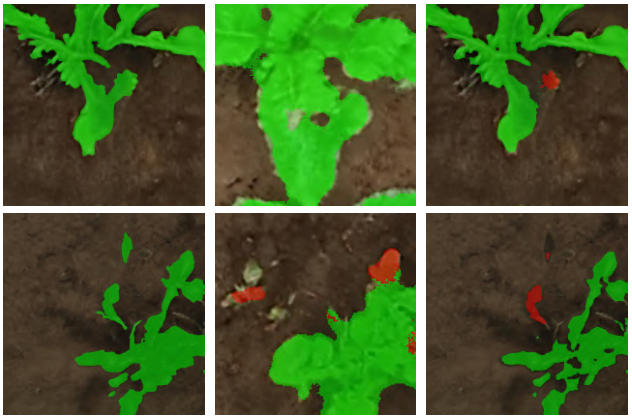
IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

TABLE III: Ablation study for performance of crop-weed segmentation on target domain (UGV domain). **Best** and second best performances are in bold and underlined respectively.

Training dataset	mIoU / %			
	non-vegetation	crop	weed	average
style inversion only	99.3	**61.8**	15.0	58.7
manual classification	99.2	59.6	**28.3**	**62.4**
sampled style inversion (**Ours**)	**99.3**	61.7	25.9	62.3

TABLE IV: mIoU of generated labels against manual labels. With label refinement, we report a higher average mIoU.

Training dataset	mIoU / %			
	non-vegetation	crop	weed	average
Without label refinement	83.7	15.3	**30.9**	43.3
With label refinement (**Ours**)	**98.1**	**44.8**	21.8	**54.9**



(a) $\hat{\mathbf{I}}_{S\to T}$ manually labeled (b) $\hat{\mathbf{Y}}_S$ on $\hat{\mathbf{I}}_S$ (c) $\hat{\mathbf{Y}}_{S\to T}$ on $\hat{\mathbf{I}}_{S\to T}$ (Ours)

Fig. 7: Qualitative results of label fitting. We indicate labeled images by overlaying the image with crops for green and weeds for red. We refined $\hat{\mathbf{Y}}_S$ to $\hat{\mathbf{Y}}_{S\to T}$, which better matches $\hat{\mathbf{I}}_{S\to T}$.

crops, while our approach does not suffer from such artifacts. Moreover, since the approach proposed by Gogoll et al. [12] conditions the contents of the generated image to be the same as that of the UAV image, the variety in the UAV images constrains the diversity of the images generated. Meanwhile, since we sample the latent space distribution to generate our images, we can generate a more diverse set of images matching the appearance of the target domain.

In summary, our approach generates images which is representative of the target domain.

D. Ablation Study

As an ablation study, we report the performance of our approach with different variations. To show the impact of sampling of the latent space during the generation of images, we report the results of two alternative approaches on the downstream task described in Sec. IV-B. We report the performance of the approach where we directly use the outputs of F_{EN}, $\mathbf{w}_S$ and $\mathbf{w}_T$, i.e., the results of the style inversion, for style mixing. We also report the performance of the approach where we manually separate UAV and UGV images from images generated by randomly sampling the latent space. Tab. III shows the results of these approaches.

Our results show that the approach with manual classification performs the best on average over all classes. The better performance may be attributed to the higher diversity of generated images since we sample the entire latent space. However, this approach requires human intervention. In contrast, the approach with style inversion only performs poorly on weeds. While this approach is relatively straightforward compared to the other approaches, we find that the diversity of the UAV dataset constrains the diversity of the generated images, which may explain the poorer performance for the weed class. We choose the sampling of the latent space approach because it does not need manual effort and performs comparably with the manual classification approach.

To show the effect of the label refinement, we report how well the generated labels fit the generated images with and without label refinement in Tab. IV. To evaluate this, we manually label a subset of 10 generated images as ground truth and calculate the mIoU of the generated labels.

We show qualitative comparison in Fig. 7. Overall, our label refinement improves the mIoU of the generated labels. Fig. 7 supports the quantitative results, where our label refinement better follows the boundaries of the crops. However, our label refinement does decrease the mIoU of weeds, where our approach mislabels crops as weeds.

V. Conclusion

In this paper, we presented a novel approach to improve the performance of a crop-weed segmentation network for agricultural robots in the target domain with a new field and on a different robotic platform. Our approach exploits StyleGAN2's style mixing properties to generate diverse target domain images in an unsupervised manner. We also proposed an unsupervised method that generates target domain image-label pairs for training the crop-weed segmentation network, thereby improving the network's performance in the target domain. We implemented and evaluated our approach on UAV and UGV image datasets and provided comparisons to existing techniques. The experiments suggest that our approach is suitable for improving the performance of the crop-weed segmentation network on new fields with a UGV, given labeled UAV images from an old field.

Despite the encouraging results, there is space for future improvement. We will explore the adaptability of our approach for larger domain gaps, e.g., between different crop types. In this paper, we assumed that sufficient amount of unlabeled target domain data is available to train our GAN, but in the future we will look into scenarios with limited amount of data.

Our approach reduces the performance degradation in new domains without the need for extra manually labeled data. This is a step towards improving the feasibility of deploying machine learning for robotics in the agricultural industry for more efficient and sustainable solutions in the near future.

REFERENCES

[1] A. Ahmadi, M. Halstead, and C. McCool. BonnBot-I: A Precise Weed Management and Crop Monitoring Platform. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2022.

[2] R. Aljundi, R. Emonet, D. Muselet, and M. Sebban. Landmarks-Based Kernelized Subspace Alignment for Unsupervised Domain Adaptation. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[3] N. Chebrolu, P. Lottes, T. Laebe, and C. Stachniss. Robot Localization Based on Aerial Images for Precision Agriculture Tasks in Crop Fields. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2019.

[4] N. Chebrolu, P. Lottes, A. Schaefer, W. Winterhalter, W. Burgard, and C. Stachniss. Agricultural Robot Dataset for Plant Classification, Localization and Mapping on Sugar Beet Fields. *Intl. Journal of Robotics Research (IJRR)*, 36(10):1045–1052, 2017.

[5] A. Chen, R. Liu, L. Xie, Z. Chen, H. Su, and J. Yu. SofGAN: A Portrait Image Generator with Dynamic Styling. *ACM Trans. on Graphics (TOG)*, 41(1):1–26, 2022.

[6] Y. Chen, Y. Lin, M. Yang, and J. Huang. CrDoCo: Pixel-Level Domain Transfer With Cross-Domain Consistency. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[7] M.J. Chong and D. Forsyth. JoJoGAN: One Shot Face Stylization. In *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2022.

[8] L. Dischinger, M. Cravetz, J. Dawes, C. Votzke, C. VanAtter, M. Johnston, C. Grimm, and J. Davidson. Towards Intelligent Fruit Picking with In-Hand Sensing. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2021.

[9] T. Duckett, S. Pearson, S. Blackmore, B. Grieve, W. Chen, G. Cielniak, J. Cleaversmith, J. Dai, S. Davis, C. Fox, P. From, I. Georgilas, R. Gill, I. Gould, M. Hanheide, A. Hunter, F. Iida, L. Mihalyova, S. Nefti-Meziani, G. Neumann, P. Paoletti, T. Pridmore, D. Ross, M. Smith, M. Stoelen, M. Swainson, S. Wane, P. Wilson, I. Wright, and G. Yang. Agricultural Robotics: The Future of Robotic Agriculture. *arXiv preprint*, arvix:1806.06762v2, 2018.

[10] M. Fawakherji, C. Potena, I. Prevedello, A. Pretto, D.D. Bloisi, and D. Nardi. Data Augmentation Using GANs for Crop/Weed Segmentation in Precision Farming. In *Proc. of the IEEE Conf. on Control Technology and Applications (CCTA)*, 2020.

[11] W. Förstner and B. Wrobel. *Photogrammetric Computer Vision*, chapter Geometry and Orientation of the Single Image, pages 457–458. Springer Verlag, 2016.

[12] 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.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks. In *Proc. of the Conf. on Neural Information Processing Systems (NIPS)*, 2014.

[14] J. Hoffman, E. Tzeng, T. Park, T. Zhu, P. Isola, K. Saenko, A.A. Efros, and T. Darrell. CyCADA: Cycle-Consistent Adversarial Domain Adaptation. In *Proc. of the Intl. Conf. on Machine Learning (ICML)*, 2018.

[15] O. Kafri, O. Patashnik, Y. Alaluf, and D. Cohen-Or. StyleFusion: Disentangling Spatial Segments in StyleGAN-Generated Images. *ACM Trans. on Graphics*, 41(5):1–15, 2022.

[16] M. Kang, J. Shin, and J. Park. StudioGAN: A Taxonomy and Benchmark of GANs for Image Synthesis. *arXiv preprint*, arXiv:2206.09479, 2022.

[17] T. Karras, S. Laine, and T. Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[18] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and Improving the Image Quality of StyleGAN. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[19] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila. Improved precision and recall metric for assessing generative models. In *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*, 2019.

[20] C. Lanczos. *Applied Analysis*. Courier Corporation, 1988.

[21] R. Li, W. Cao, S. Wu, and H.S. Wong. Generating Target Image-Label Pairs for Unsupervised Domain Adaptation. *IEEE Transactions on Image Processing*, 29:7997–8011, 2020.

[22] S. Lo, W. Wang, J. Thomas, J. Zheng, V. Patel, and C. Kuo. Learning Feature Decomposition for Domain Adaptive Monocular Depth Estimation. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2022.

[23] 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(4):3097–3104, 2018.

[24] P. Lottes. *Plant Classification Systems for Agricultural Robots*. PhD thesis, Rheinische Friedrich-Wilhelms University of Bonn, 2021.

[25] E. Marks, F. Magistri, and C. Stachniss. Precise 3D Reconstruction of Plants from UAV Imagery Combining Bundle Adjustment and Template Matching. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2022.

[26] M.F. Naeem, S.J. Oh, Y. Uh, Y. Choi, and J. Yoo. Reliable fidelity and diversity metrics for generative models. In *Proc. of the Intl. Conf. on Machine Learning (ICML)*, 2020.

[27] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*, 2019.

[28] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or. Encoding in Style: A StyleGAN Encoder for Image-to-Image Translation. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[29] G. Roggiolani, M. Sodano, T. Guadagnino, F. Magistri, J. Behley, and C. Stachniss. Hierarchical Approach for Joint Semantic, Plant Instance, and Leaf Instance Segmentation in the Agricultural Domain. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2023.

[30] 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 (T-ITS)*, 19(1):263–272, 2018.

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *Intl. Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[32] C. Smitt, M. Halstead, T. Zaenker, M. Bennewitz, and C. McCool. PATHoBot: A robot for glasshouse crop phenotyping and intervention. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2021.

[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception Architecture for Computer Vision. *arXiv preprint*, arxiv:1512.00567v3, 2015.

[34] A. Tewari, M. Elgharib, G. Bharaj, F. Bernard, H. Seidel, P. Perez, M. Zollhofer, and C. Theobalt. StyleRig: Rigging StyleGAN for 3D Control Over Portrait Images. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[35] S.G. Vougioukas. Agricultural robotics. *Annual Review of Control, Robotics, and Autonomous Systems*, 2(1):365–392, 2019.

[36] J. Wang, W. Feng, Y. Chen, H. Yu, M. Huang, and P.S. Yu. Visual Domain Adaptation with Manifold Embedded Distribution Alignment. In *Proc. of the ACM Intl. Conf. on Multimedia*, 2018.

[37] W. Wang, H. Zhang, Z. Yuan, and C. Wang. Unsupervised Real-World Super-Resolution: A Domain Adaptation Perspective. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2021.

[38] J. Weyler, F. Magistri, P. Seitz, J. Behley, and C. Stachniss. In-Field Phenotyping Based on Crop Leaf and Plant Instance Segmentation. In *Proc. of the IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2022.

[39] J. Weyler, J. Quakernack, P. Lottes, J. Behley, and C. Stachniss. Joint Plant and Leaf Instance Segmentation on Field-Scale UAV Imagery. *IEEE Robotics and Automation Letters (RA-L)*, 7(2):3787–3794, 2022.

[40] X. Wu, S. Aravecchia, P. Lottes, C. Stachniss, and C. Pradalier. Robotic weed control using automated weed and crop classification. *Journal of Field Robotics (JFR)*, 37:322–340, 2020.

[41] W. Xu, Y. Li, and C. Lu. SRDA: Generating Instance Segmentation Annotation via Scanning, Reasoning and Domain Adaptation. In *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2018.

[42] X. Xu, P. Wei, W. Chen, Y. Liu, M. Mao, L. Lin, and G. Li. Dual Adversarial Adaptation for Cross-Device Real-World Image Super-Resolution. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[43] J. Zhu, T. Park, P. Isola, and A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2017.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.