Zero-Shot Semantic Segmentation for Robots in Agriculture

Yue Linn Chong Lucas Nunes Federico Magistri Xingguang Zhong Jens Behley Cyrill Stachniss

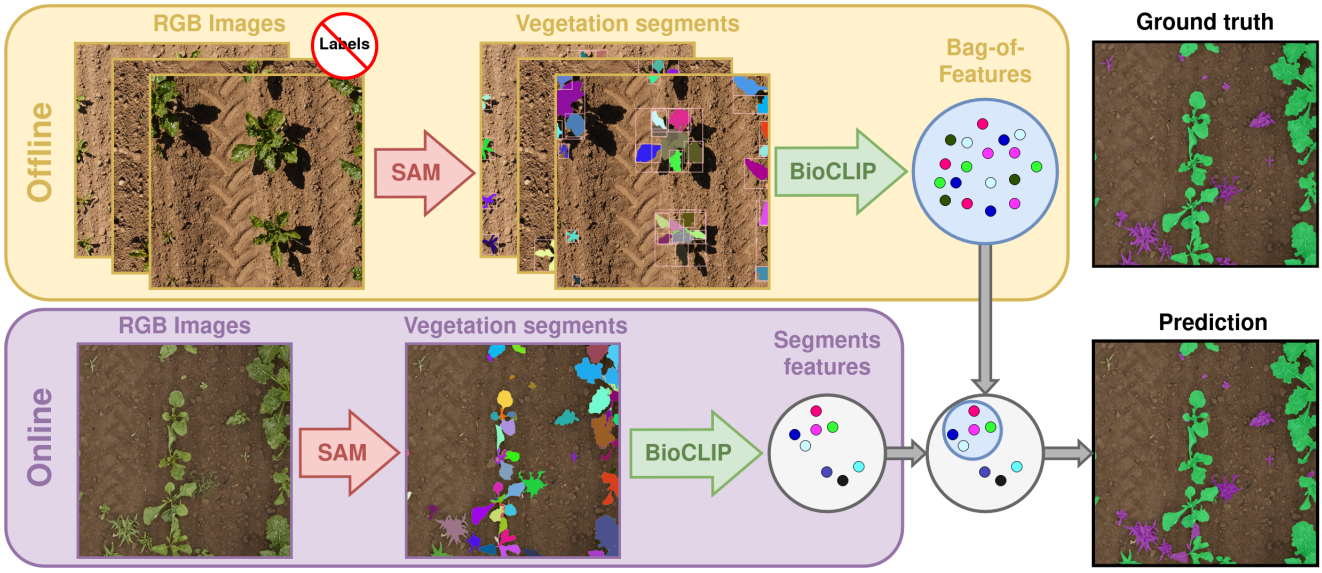


Fig. 1: Our approach can segment crop plants and weeds without labels. We leverage foundation models SAM [17] and BioCLIP [34] to build a bag of features representing crop plants. During inference, we extract plant features and compare them with the bag of features. Plant features with low similarity to the bag of features are inferred as weeds.

***Abstract*— Conventional crop production, which is essential for providing food, feed, fuel, and fiber for our society, relies heavily on harmful herbicides to control weeds. Instead, agricultural robots could remove weeds more sustainably. However, these robots require a generalizable perception system that can locate weeds, enabling automatic removal of weeds. Specifically, they need to perform crop-weed semantic segmentation, which locates and distinguishes between the crop and the weed plants with pixel-level resolution. However, most existing crop-weed semantic segmentation methods are fully supervised and require expensive and labor-intensive pixel-wise labeling of the training data. To avoid the costly labeling process, we address the problem of unsupervised crop-weed segmentation in this paper. Unlike previous approaches, we leverage the idea that weeds are "weird" plants that occur less frequently and are highly variable in appearance, and reframe the problem as an anomaly segmentation problem. We propose an approach to segment weeds as anomalous plants by categorizing plants in the feature space of a pretrained foundation model. Our approach curates a bag-of-features representation of crop features and models the manifold of crop plants as hyperspheres. During inference, it classifies vegetation segments of the image with features within this manifold as crop plants and all other plants as weeds. Our experiments show that our zero-shot anomaly segmentation method can perform crop-weed segmentation on several datasets from real crop fields.**

All authors are with the Center for Robotics, University of Bonn, Germany. Cyrill Stachniss is additionally 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.

I. Introduction

Current agricultural practices uniformly spray herbicides on the field to combat weeds. This non-targeted application of herbicides harms our environment and its biodiversity. Precision farming robots can reduce herbicide usage without compromising yield by applying herbicides only to selected areas or specific individual weed plants or with mechanical weeding [7], [24]. The first step towards automated weeding with agricultural robots is to locate plants and distinguish between crops and weeds in RGB images captured of the field. To estimate the amount of required herbicide and precise application, we require the localization of the weeds to be at pixel-level, i.e., semantic segmentation of weed and crop. While fully-supervised methods [20], [39] can perform crop-weed segmentation well, they require in-domain labels for training. Obtaining these labels can be very costly since pixel-wise manual labeling is time-consuming [39]. Additionally, agricultural applications encompass diverse domains, with diverse soil and lighting conditions, different robot platforms, i.e., unmanned aerial vehicles (UAVs) and unmanned ground vehicles (UGVs), and various crop appearances, and domain shifts degrade models' performance [36]. Scaling manual labeling to cover the diverse agricultural domains is not cost-effective. Thus, in this paper, we aim to avoid relying on fully supervised methods for automatic weeding and move towards a zero-shot setting, thereby eliminating the need for additional labeling requirements.

The main contribution of this paper is twofold: firstly, we propose a zero-shot approach for crop-weed semantic segmentation by identifying the crops as the most commonly occurring plant and the weeds as "weird"-looking plants. Secondly, our approach is also able to perform zero-shot vegetation segmentation that generalizes well across multiple datasets, by leveraging the segment anything model (SAM) [17]. Fig. 1 shows the overview of our approach. Our code is available at `https://github.com/PRBonn/WeedsAreWeird`.

In summary, we make these key claims: Our approach is able to (i) perform crop-weed semantic segmentation on real-world UAV and UGV datasets without additional labels, and (ii) perform vegetation segmentation using our novel prompting method, which generalizes well across multiple UAV and UGV datasets. These claims are backed up by our experimental evaluation in Sec. IV.

II. Related Work

A. Semantic Segmentation

For crop-weed semantic segmentation, fully supervised approaches using convolutional neural networks (CNNs) perform well [20], [39] and are the de facto standard solution for most robotic vision tasks in crop fields [42], [43]. However, these fully supervised methods require pixel-wise annotation, which is labor-intensive to acquire. Several works proposed different approaches to reduce the amount of labeling effort required. Some works leverage the use of geometric heuristics, specifically utilizing the crop-row structure to automatically identify crops [20], [38]. However, these geometric structures are not necessarily present in all scenarios and are particularly lacking where the camera field of view is small [2]. Moreover, some weeds grow within crop rows, which breaks this heuristic. Different research directions involve the use of unsupervised domain adaptation approaches [6], [10], [22] or training networks with only partial labels [37], but these directions, albeit showing promising results, still require labeled images. In contrast, our approach does not require any in-domain labeling and instead leverages pretrained foundation models. In line with existing methods that leverage foundation models [19], [29], we refer to our approach as a zero-shot method.

Similar to prior work by Fawakherji et al. [12], our modular approach performs intermediate vegetation segmentation by classifying each pixel as plant or soil. There are many heuristic vegetation segmentation algorithms [4], [13], [16], [21], [23], [25], [26], [35], [40]. While segmenting plants may sound trivial, vegetation segmentation can be challenging due to the diversity of plants and environmental conditions such as lighting and soil texture. At the very least, it requires tuning new hyperparameters for each domain, which causes most work to fail in previously unseen domains.

In our work, we present a domain-generalized zero-shot method leveraging SAM [17]. Unlike the work by Carraro et al. [8], which also utilizes SAM for vegetation annotation, we employ a different sampling and filtering method to overcome the over-segmentation of soil. Moreover, our zero-shot semantic segmentation approach differs from existing work [29], which also leverages SAM [17] and CLIP [28]. In particular, existing work Grounded SAM [29] relies on text prompting, but text prompts are ill-defined in the crop-weed segmentation task. Instead, our approach leverages information from the dataset in its entirety to define common plants as crops and weird plants as weeds.

B. Unsupervised Anomaly Segmentation

Our work is closely related to approaches used for unsupervised anomaly detection and segmentation, typical of agriculture inspection [1], [9], industrial inspection [11], and medical imaging [5], [41]. However, these use cases have few examples of anomalies but many examples of normal data, i.e., data without anomalies present. The anomalies are also diverse and may differ greatly in visual appearance. These two conditions are similar to our use case for crop-weed segmentation. Normalities are synonymous with our crops, and the weeds are the anomalies that are visually diverse and fewer in number. While the paradigm is considered unsupervised, such methods are actually weakly supervised since a set of images had to be classified as normal for training.

There are two broad categories of unsupervised anomaly segmentation methods: (1) feature-based filtering, (2) generative modeling, and (3) semantic segmentation-adjacent methods. Methods from the first category, such as PaDiM [11], maintain a bag of features from normal images, with features obtained using a pre-trained encoder. During testing, we detect the anomalies by identifying features that are dissimilar to the known normal bag of features. Our approach falls into this category.

Methods from the second category [3], [5], [41] learn the distribution of the normal data using a generative model to conditionally generate normal data. At test time, regions with high reconstruction error are predicted as anomalous regions since the methods were not trained to generate anomalies.

While there are many similarities between our work and unsupervised anomaly segmentation approaches, there are two key differences. Firstly, weeds are present in the dataset, and we cannot easily curate a dataset without weeds. We propose that these weeds are noise in the training data. There are a few unsupervised anomaly segmentation works that account for noisy training data. SoftPatch [15] addresses noisy anomaly segmentation training data by weighing down anomalous features in the training data. Secondly, data from the agriculture domain is more complex than other anomaly segmentation datasets because of the relatively uncontrolled real-world crop field, and the location of the plants occurring anywhere in the image.

We propose our method with these differences in mind. Similar to SoftPatch [15], we first extract features from all images in the noisy dataset. However, unlike SoftPatch, we remove anomalies (or weeds) by choosing the most commonly occurring features (of the crop plants) instead of eliminating uncommon features. Finding the crop plant

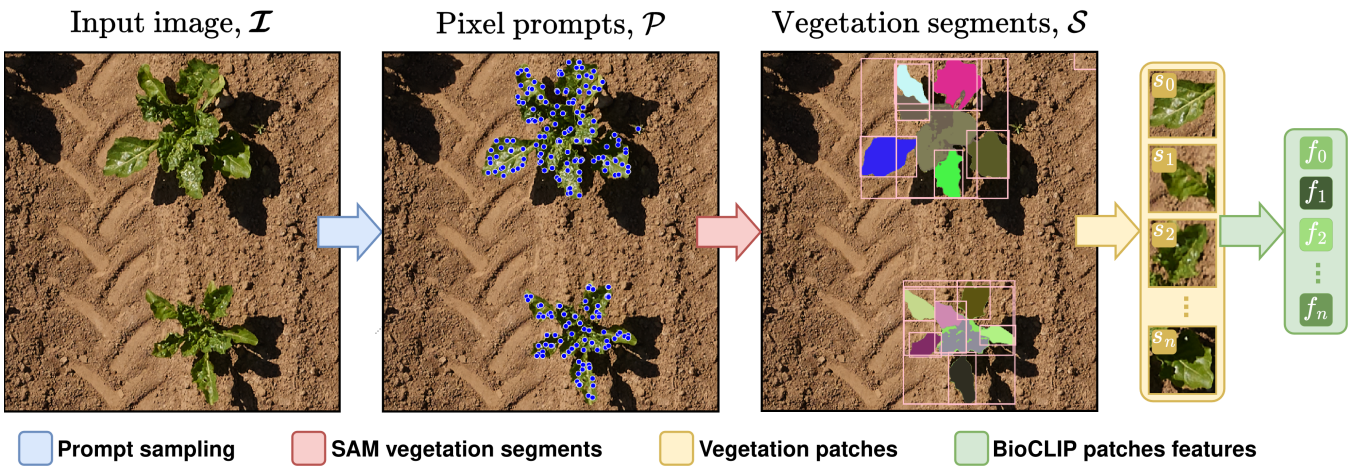


Fig. 2: Overview of how we extract features, $\mathbf{f}_i, i = \{0, 1, \ldots, n\}$, from an image, $\mathcal{I}$. We identify point prompts, $\boldsymbol{p}_i$, using ExG. With each $\boldsymbol{p}_i$, we obtain segment s_i using SAM. We input the cropped patches, $\hat{\mathcal{I}}_i$, of s_i to BioCLIP to obtain the patch feature $\mathbf{f}_i$.

features this way is easier for our use case since there are more anomalies (i.e., weeds) in our dataset compared to that of SoftPatch. Note that SoftPatch was developed for anomaly detection in the structured scenario of manufacturing defects, whereas our method focuses on crop-weed semantic segmentation.

III. Our Approach

We perform zero-shot crop-weed semantic segmentation by leveraging the foundation models SAM [17] and BioCLIP [34]. Fig. 1 provides the overview of our approach, called Weeds are Weird (WaW). Our approach has two steps. The first is to curate a bag of features $\mathcal{B}$ representing the crop plants, as explained in Sec. III-A. The second step is to perform inference using $\mathcal{B}$, as elaborated in Sec. III-B.

A. Curating the Bag of Features

Fig. 2 shows an overview of how we obtain features from a given RGB image $\mathcal{I}$. From $\mathcal{I}$, we create a set of point prompts $\mathcal{P}$ to prompt SAM [17], resulting in vegetation segments $\mathcal{S}$. With the bounding boxes of segments $s \in \mathcal{S}$, we crop $\mathcal{I}$ to obtain cropped patches $\hat{\mathcal{I}}$. The patches $\hat{\mathcal{I}}$ are input to BioCLIP [34] to obtain patch features $\mathbf{f}_i \in \mathbb{R}^{512}, i = 1, 2, \ldots, |\hat{\mathcal{I}}|$. We repeat this process with all images in the training set to form a set of features $\mathcal{F}$. From $\mathcal{F}$, we use the popularity voting algorithm, formalized in Alg. 1, to obtain the bag of features $\mathcal{B}$ representing the crop features used later to identify weeds.

1) Vegetation Segments: We begin by predicting the vegetation segments $\mathcal{S}$ in image $\mathcal{I}$. At this stage, we aim to locate the plants in $\mathcal{I}$ without needing to distinguish between crop plants and weeds. The segments do not necessarily need to be complete plant instances (where the instance comprises all the pixels representing an individual plant) and can be segments representing partial instances.

We obtain the segments using SAM [17] with point prompts $\boldsymbol{p} \in \mathcal{P}$, where $\boldsymbol{p} \in \mathbb{R}^2$. For generating the point prompts $\mathcal{P}$, we subdivide the image into $N \times N$ grid cells and generate for each grid cell a point prompt $\boldsymbol{p}$ for the pixel location, (x, y), with the highest excess green vegetation index (ExG) [40] score, if $\text{ExG}(x, y) > \lambda_{exg}$. We use each $\boldsymbol{p} \in \mathcal{P}$ individually to prompt SAM without multi-masks to return a single segment s_i with score $\hat{s}_i$ per prompt, $\boldsymbol{p}$, i.e.,

$$\mathcal{S}' = \{(s_i, \hat{s}_i) = \text{SAM}(\boldsymbol{p}) \mid \boldsymbol{p} \in \mathcal{P}\}. \quad (1)$$

First, we remove all segments $(s_i, \hat{s}_i) \in \mathcal{S}'$, where most of the points are most likely soil components, i.e., if:

$$\frac{|\{(x, y) \in s_i \mid \text{ExGR}(x, y) > \lambda_{exgr}\}|}{|s_i|} < \lambda_{\text{percent}}, \quad (2)$$

where, $|s_i|$ corresponds to the number of pixels in segment s_i, and $\text{ExGR}(x, y)$ returns the excess green minus excess red (ExGR) index [26] at pixel location (x, y).

We refine $\mathcal{S}'$ using a non-maximum suppression, where the confidence c_i of the segment $(s_i, \hat{s}_i) \in \mathcal{S}'$ is given by:

$$c_i = \hat{s}_i \frac{|\{(x, y) \in s_i \mid \text{ExG}(x, y) > \lambda_{exg}\}|}{|s_i|}, \quad (3)$$

where $\text{ExG}(x, y)$ returns the ExG vegetation index at pixel location (x, y). If segments $s_i \in \mathcal{S}'$ and $s_j \in \mathcal{S}'$ have an

Algorithm 1 Popularity Voting Algorithm

Input: $\mathcal{F}$
Output: $\mathcal{B}$
for iteration=1,2,...,M **do**
  $\hat{\mathcal{F}}_{\text{candidate}} \leftarrow$ sample $N_{\text{candidate}}$ features from $\mathcal{F}$
  $\hat{\mathcal{F}}_{\text{population}} \leftarrow$ sample $N_{\text{population}}$ features from $\mathcal{F}$
  for $\mathbf{f}_{\text{cand,i}} \in \hat{\mathcal{F}}_{\text{candidate}}$ **do**
    for $\mathbf{f}_{\text{pop,j}} \in \hat{\mathcal{F}}_{\text{population}}$ **do**
      $w_{i,j} \leftarrow w_{\text{size}}(\text{size}_i, \text{size}_j)$
      $\text{sim}_{i,j} \leftarrow w_{i,j} \cos\_\text{sim}(\mathbf{f}_{\text{cand},i}, \mathbf{f}_{\text{pop},j})$
  $i \leftarrow \underset{i}{\text{argmax}}(\text{sim}_j)$
  $\mathcal{B} \leftarrow \mathcal{B} \cup \hat{\mathbf{f}}_{\text{cand},i}$
  $\mathcal{F} \leftarrow \mathcal{F} - \hat{\mathbf{f}}_{\text{cand},i}$
return $\mathcal{B}$

intersection-over-union (IoU) of at least λ_{NMS}, we only keep the segment s_i with the higher confidence c_i, which results in the final set of vegetation segments $\mathcal{S}$.

2) Feature Encoding: We used the pre-trained BioCLIP [34] to extract features for each segment $s_i \in \mathcal{S}$ to obtain $\mathbf{f}_i \in \mathbb{R}^{512}$ since our task requires the distinction between plant species and BioCLIP is trained for fine-grained biodiversity classification.

For each vegetation segment $s_i \in \mathcal{S}$, we crop out a patch $\hat{\boldsymbol{\mathcal{I}}}_i$ given by the square bounding box of s_i from image $\boldsymbol{\mathcal{I}}$, where $\hat{\boldsymbol{\mathcal{I}}}_i$ is the crop patch with side length $\max(w, h)$ given the bounding box of s_i of width w and height h. We input $\hat{\boldsymbol{\mathcal{I}}}_i$ to BioCLIP to obtain the patch features $\mathbf{f}_i$, i.e.,

$$\mathbf{f}_i = \text{BioCLIP}(\text{resize}(\hat{\boldsymbol{\mathcal{I}}})), \tag{4}$$

where $\text{resize}(\cdot)$ bilinearly rescales an image to 224×224 px. We perform feature encoding on all vegetation segments from training images and obtain a pool of features $\mathcal{F}$, containing features of crop plants and weeds. We select the crop features only using our popularity voting algorithm, shown in Alg. 1, where $\cos\_\text{sim}(\cdot, \cdot)$ is the cosine similarity between two input vectors. The key idea is that crop plant features have a high occurrence frequency and, therefore, have the highest similarity to most of the other features.

In each iteration, we form a subset $\hat{\mathcal{F}}_{\text{candidate}} \subset \mathcal{F}$, where $|\hat{\mathcal{F}}_{\text{candidate}}| = N_{\text{candidate}}$, as candidate crop plant features. As the number of features in $\mathcal{F}$ is large, we randomly subsample features into a smaller representative set, $\hat{\mathcal{F}}_{\text{population}}$, where $|\hat{\mathcal{F}}_{\text{population}}| = N_{\text{population}}$. We evaluate the similarity score sim_i of each $\hat{\mathbf{f}}_{\text{cand},i}$ as the sum of the cosine similarity between $\hat{\mathbf{f}}_{\text{cand},i}$ and each feature in $\hat{\mathbf{f}}_{\text{pop,j}}, j = 1, \ldots, \lambda_{\text{population}}$. The similarity score is weighted with $w_{\text{size}}(\text{size}_i, \text{size}_j)$, given by:

$$w_{\text{size}}(\text{size}_i, \text{size}_j) = \cos\_\text{sim}(g(\text{size}_i), g(\text{size}_j)), \tag{5}$$

with

$$g(x) = \mathcal{N}(0.0, 0.7) * (\mathbb{I}(x, 0), \ldots, \mathbb{I}(x, K-1)), \tag{6}$$

where $*$ is the convolution operation, and

$$\mathbb{I}(x, i) = \begin{cases} 1 & , \lfloor (x - \text{size}_{\text{min}})/\delta_{\text{size}} \rfloor = i \\ 0 & , \text{otherwise} \end{cases}. \tag{7}$$

Here, K is the number of size categories, size_{min} is the smallest patch size category, size_{max} is the largest patch size category, size_i is the size of patch $\hat{\boldsymbol{\mathcal{I}}}_i$, δ_{size} is the step size for the class patch size, where

$$\delta_{\text{size}} = (\text{size}_{\text{max}} - \text{size}_{\text{min}})/K, \tag{8}$$

and $\mathcal{N}(\mu, \sigma)$ is a 1-dimension Gaussian kernel with mean μ and standard deviation σ. The weighting $w_{\text{size}}(\text{size}_i, \text{size}_j)$ increases when $\hat{\boldsymbol{\mathcal{I}}}_i$ and $\hat{\boldsymbol{\mathcal{I}}}_j$ are similar in size and gradually reduces to zero when $\hat{\boldsymbol{\mathcal{I}}}_i$ and $\hat{\boldsymbol{\mathcal{I}}}_j$ grow increasingly dissimilar in size. This weighting is required because we resized all patches to the same size in preprocessing for BioCLIP. Each $\hat{\mathbf{f}}_{\text{pop,j}}$ can only vote for $\hat{\mathbf{f}}_{\text{cand,i}}$. We add the candidate feature with the highest total of voting scores to $\mathcal{B}$. We repeat the voting procedure M times to form the bag of features $\mathcal{B}$.

TABLE I: Hyperparameters used for each dataset.

Dataset	k	M
PhenoBench [39]	10	500
SB20 [2]	200	1000
CropAndWeed-SugarBeet [33]	20	500
CropAndWeed-Maize [33]	50	500

B. Inference

To perform inference, we repeat the vegetation segmentation and feature encoding as shown in Fig. 2 to collect a set of features $\mathcal{F}_{\text{infer}}$ of the input image. For each feature $\mathbf{f}_{\text{infer}} \in \mathcal{F}_{\text{infer}}$, we check if $\mathbf{f}_{\text{infer}}$ is within the manifold of $\mathcal{B}$. We model the manifold by the union of hyperspheres, similar to that used by Kynkäänniemi et al. [18]. Specifically, we model the manifold of crop plants using hyperspheres h_i with centers at features $\hat{\mathbf{f}}_i \in \mathcal{B}$. The radius of h_i is equivalent to the Euclidean distance of $\hat{\mathbf{f}}_i$ to its k-th nearest neighbor. If $\mathbf{f}_{\text{infer}}$ is within the manifold of $\mathcal{B}$, we classify pixels in s_{infer} as a crop plant. Otherwise, we classify all pixels in s_{infer} as weed. Since some segments, s_i, overlap, we take the average of each s_i for the final classification. Finally, we classify all the remaining pixels as soil.

IV. Experiments and Discussions

The main focus of this work is an approach for zero-shot crop-weed semantic segmentation for agricultural robots, without additional labels, by framing weeds as anomalies. To show the capabilities of our approach, we conduct experiments for two main applications in the agricultural domain: crop-weed segmentation and vegetation segmentation. The experiments show that our method can adapt to various environmental conditions.

A. Experimental Setup

Datasets. We selected multiple datasets providing annotations for crops and weeds that cover different domains and crops: PhenoBench [39], SB20 [2], and CropAndWeed [33]. PhenoBench [39] is a labeled dataset of high-resolution UAV images of a sugar beet field, and SB20 [2] comprises labeled images from a UGV of a sugar beet field. The CropAndWeed [33] dataset comprises images from multiple fields of various crops. We split the CropAndWeed dataset into separate single-crop datasets to avoid identifying volunteer crops as crops, and used the data for sugar beets and maize only. While the datasets are labeled, our approach and baselines do not use the labels for training.

For variation in crop cultivar, we included a dataset with maize crops from the CropAndWeeds dataset [33], whereas the other datasets are of sugar beet crops. For the selection of datasets, we also considered the number of available images, as our baselines, AnoDDPM [41] and THOR [5], require training diffusion models, which likely perform better with a larger number of images. For our approach, we populate the bag of features $\mathcal{B}$ with images from the training split, and for SB20, we also used the provided unlabeled images.

Implementation Details. For all our experiments, we used $N = 48$, $\lambda_{exg} = 0.0$, $\lambda_{exgr} = 0.0$, $\lambda_{NMS} = 0.7$,

TABLE II: IoU Performance on test sets %. **Bold** indicates the best performance.

Method	PhenoBench [39]				SB20 [2]				CropAndWeed-SugarBeets [33]				CropAndWeed-Maize [33]				Mean	Std.
	soil	crop	weed	mIoU	soil	crop	weed	mIoU	soil	crop	weed	mIoU	soil	crop	weed	mIoU	weed IoU	
ERFNet [30]	99.3	94.3	64.4	86.0	98.4	79.0	72.3	83.3	99.3	88.6	54.8	80.9	99.2	77.1	59.2	78.5	62.7	7.5
AnoDDPM [41]	**99.1**	86.0	3.9	63.0	**97.1**	34.2	7.0	46.1	98.0	26.6	9.6	44.8	97.5	36.6	7.9	47.3	7.1	2.4
THOR [5]	**99.1**	**88.4**	4.0	**63.8**	**97.1**	**50.3**	0.4	**49.2**	98.0	**57.1**	1.2	52.1	97.5	**55.0**	1.3	51.3	1.7	**1.6**
WaW (Ours)	98.9	74.3	**11.1**	61.4	96.6	29.5	**17.1**	47.7	**98.6**	52.7	**9.9**	**53.7**	**99.0**	46.2	**10.2**	**51.8**	**12.1**	3.4

$\lambda_{\text{percent}} = 0.2$, $K = 5$, $\text{size}_{\text{min}} = 0$, $\text{size}_{\text{max}} = 200$, $N_{\text{population}} = 100$, and $N_{\text{candidate}} = 100$. We qualitatively tuned the hyperparameters k and M for each dataset, as these hyperparameters depend on $|\mathcal{S}|$. See Tab. I for the dataset-specific values. On average, our inference runtime on an NVIDIA RTX A6000 GPU with batch size one is 5.6 s.

Metrics. We calculate the intersection-over-union (IoU) for each class (i.e., soil, crop, and weed) and also compute the mean over all classes to obtain the mIoU, similar to previous crop-weed semantic segmentation works [39].

B. Performance on Crop-Weed Semantic Segmentation

The first experiment evaluates the performance of our approach in crop-weed semantic segmentation with no additional manual labels. Since our approach falls under the paradigm of unsupervised anomaly segmentation, we adapted approaches from this field for comparison. We compare our approach with other unsupervised baselines: AnoDDPM [41] and THOR [5]. We trained the diffusion model used in both baselines for 255,000 steps for all datasets. Since these methods only output anomaly segments, we convert the anomaly masks to crop-weed semantic segmentation using the ExG [40] vegetation mask. We classify pixels that are vegetation and anomalies as weeds, and pixels that are vegetation but not anomalies as crops. For an upper-bound comparison, we also report the performance of the fully supervised method ERFNet [30], which shows promising results in crop-weed segmentation [39].

Tab. II shows the performance of our method and the baselines and Fig. 3 illustrates the qualitative results of our approach. The results show that the unsupervised anomaly segmentation methods have the potential to perform zero-shot crop-weed semantic segmentation. Since our downstream task involves automatic weeding, we focus on the IoU of weeds, which is also more challenging to segment compared to crops, as indicated by the consistently lower IoU of weeds compared to that of crops. Our zero-shot method outperforms baseline methods in all datasets, despite requiring no additional training, whereas the baselines require long training times typical of diffusion models.

Unsurprisingly, our method performs poorly compared to the supervised ERFNet [30] across all metrics and datasets. This performance is to be expected, as ERFNet has access to labels, whereas our approach does not. Based on the crops and weed IoU, we see that all the unsupervised methods, including ours, have further room for improvement. Interestingly, for the IoU of the soil class, the unsupervised baselines and our approach have similar performance to that of ERFNet. The IoU of soil is constantly highest among the three classes, which indicates that the separation of soil from vegetation is simpler than that of crops and weeds. The large volume of soil pixels also further increases the average soil IoU score in general.

As shown in the qualitative results, our approach is unable to distinguish between crops and weeds in certain situations. Particularly, weeds that overlap with crops may be incorrectly classified as crops, as shown in Fig. 3(a), likely due to the patch extraction for these segments, which also includes neighboring crop plants when patched into squares for feature extraction with BioCLIP.

Since we rely on the ExG [40] index for prompting the segmentation of $\mathcal{S}$, we wrongly classify weeds that are reddish and resemble the soil color, since these reddish weeds have low vegetation index scores, as shown in Fig. 3(b).

C. Performance for Vegetation Segmentation

Our second experiment evaluates the performance of extracting vegetation masks obtained from our approach. Specifically, we evaluate the performance of our vegetation segmentation method as a semantic segmentation task with only two classes: soil and vegetation. We compare our approach with existing work on vegetation masking, where we tuned the thresholds of baselines using a subsample of training data from PhenoBench [39] and used the same threshold across all datasets.

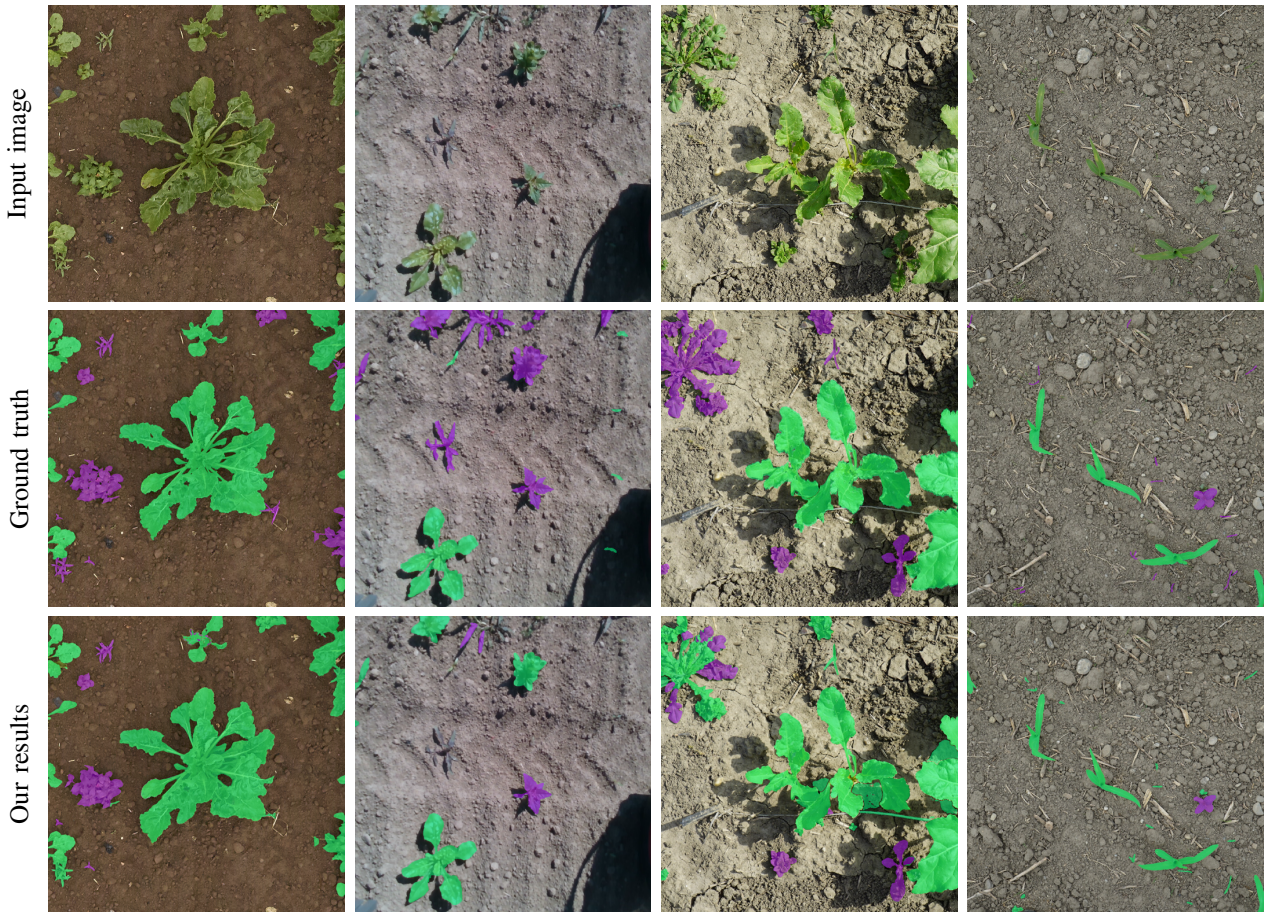
Tab. III shows the quantitative results of our approach and baselines for the vegetation segmentation task on the validation split of each dataset. Our approach outperforms the baselines in most datasets, demonstrating the benefits of leveraging segments instead of per-pixel thresholding, as used by the baseline methods. Notably, our SAM-based method has the highest vegetation IoU, which supports its use for subsequent crop-weed segmentation. Across the baseline methods, we see that some methods generalize better across different datasets. Notably, ExGR [26] and ExG [40] perform relatively well among other baselines, which supports the adaptation of these vegetation indices into our approach.

V. Ablations and Hyperparameter Search

We performed ablation studies on the prompting mechanism and feature encoder, as well as hyperparameter search for the hyperparameters used in our approach. For all studies in this section, we evaluated on a subset of 500 images from the validation split of PhenoBench [39].

A. Prompting SAM

This section discusses the impact of our proposed prompting mechanism used with SAM [17], as shown in Tab. IV. We



(a) PhenoBench [39] (b) SB20 [2] (c) CropAndWeed-SugarBeets [33] (d) CropAndWeed-Maize [33]

Fig. 3: Qualitative results on different datasets. The top row shows the input image, the second row shows the ground truth, and the third row shows our performance.

TABLE III: IoU of vegetation on validation split.

Method	PhenoBench [39]			SB20 [2]			CropAndWeed-SugarBeets [33]			CropAndWeed-Maize [33]			Mean	Std.
	soil	vegetation	mIoU	soil	vegetation	mIoU	soil	vegetation	mIoU	soil	vegetation	mIoU	mIoU	mIoU
ExG [40]	98.3	78.5	88.4	**96.2**	48.2	72.2	97.1	49.2	73.2	97.8	49.5	73.7	76.9	7.7
ExR [25]	97.8	63.9	80.8	5.7	7.3	6.5	4.4	5.4	4.9	8.9	2.7	5.8	24.5	37.6
ExGR [26]	98.4	72.3	85.4	**96.2**	49.0	72.6	98.1	53.9	76.0	98.9	54.5	76.7	77.6	5.4
GRVI [35]	98.3	68.2	83.2	44.3	11.3	27.8	22.5	6.7	14.6	30.8	5.8	18.3	36.0	32.0
MGRVI [4]	98.3	69.0	83.6	30.1	9.4	19.8	11.5	5.7	8.6	19.9	4.0	11.9	31.0	35.4
GLI [21]	98.5	78.9	88.7	95.5	38.2	66.8	97.9	51.6	74.8	98.5	52.4	75.5	76.4	9.0
RGBVI [4]	95.7	55.0	75.3	93.9	13.7	53.8	97.8	40.4	69.1	98.2	38.9	68.6	66.7	9.1
WaW (Ours)	**98.8**	**83.7**	**91.3**	95.7	**61.6**	**78.7**	**98.5**	**72.1**	**85.3**	**99.1**	**70.1**	**84.6**	**85.0**	**5.1**

TABLE IV: IoU for ablations on vegetation segment method. **Bold** indicates the best performance.

Method	Prompting Mechanism	Segment Filter	Soil IoU	Vegetation IoU	mIoU
SAM [17]	✗	✗	89.4	38.5	63.9
WaW w/ grid	✗	✓	98.0	77.4	87.7
WaW (Ours)	✓	✓	**98.9**	**84.4**	**91.7**

TABLE V: IoU for ablations on feature encoding methods. **Bold** indicates best performance.

Feature Encoder	Soil IoU	Crop IoU	Weed IoU	mIoU
ResNet-152 [14]	**98.9**	**81.3**	1.6	60.6
CLIP [27]	**98.9**	77.3	9.1	**61.8**
BioCLIP [34] (Ours)	**98.9**	73.5	**12.8**	61.7

compare our prompting mechanism with the out-of-the-box grid prompting used in SAM, ignoring the largest segment as soil. SAM performs poorly on both the soil and vegetation IoU, mostly because SAM over-segments the rocks in the soil. We also tested using a combination of the out-of-the-box SAM with just our segment filtering post-processing. This combination yielded a lower IoU, indicating the importance of the proposed prompting method for improving vegetation segmentation performance.

TABLE VI: IoU performance with varying values of k. **Bold** indicates best performance.

k	Soil IoU	Crop IoU	Weed IoU	mIoU
1	**98.9**	49.8	7.9	52.2
5	**98.9**	69.0	11.5	59.8
10	**98.9**	73.5	**12.8**	61.7
100	**98.9**	**82.0**	7.4	**62.8**
500	**98.9**	81.7	0.1	60.2

B. Importance of BioCLIP

We tested different feature encoders, replacing BioCLIP [34] used in our proposed approach with a pre-trained ResNet [14], similar to previous work [31] and the same CLIP [27] architecture, but with weights trained on a different generic (not biodiversity-specific) dataset [32]. Tab. V shows the IoU performance of these two ablations and our proposed approach. Notably, the method using ResNet features performs much worse for the weeds. While the performance when using CLIP versus BioCLIP is comparable for crop plants, the performance on the weeds is better when using BioCLIP, which supports our decision to use BioCLIP [34] in our proposed approach.

C. Hyperparameter Analysis

In this section, we discuss the impact of changing the hyperparameter values. Firstly, we investigate how changing the neighborhood size k, affects the semantic segmentation performance. Tab. VI shows the IoU performance for varying k values. We show that the highest IoU for weeds is at $k = 10$, which is the value we used in our approach. Reducing k leads to poorer performance, likely because the hyperspheres are too small, so only the features with low distance from the bag of features are predicted as crops. If k is too large, too many false positives are identified as crops, as the hyperspheres are too large, resulting in poorer semantic segmentation performance.

Secondly, we study how varying the grid size N used in our prompting mechanism affects the performance of vegetation segmentation. Tab. VII shows the results with varying values of N. Increasing N will increase the number of prompts, which leads to fewer plants not being segmented. Thus, this would increase the IoU across both soil and vegetation classes. However, increasing the number of prompts also increases the computational resources required to run the method. Hence, we chose a moderate value of $N = 48$, which corresponds to where the performance begins to plateau. Note that halving the value to $N = 24$ results in a decrease of less than 1 percentage point, which can be attributed to the prompting mechanism. With the prompting mechanism, the prompt will always land on the plant if one is present in the grid cell. Thus, fewer plants are left unprompted and therefore segmented. However, since each grid cell can contribute only one prompt, if the grid cells are too sparse, many plants remain unprompted, negatively affecting performance.

Finally, we study the impact of varying the size M of the bag of features $\mathcal{B}$. If the bag-of-features has too few features, the approach will not be able to capture the full variety of crop features. As shown in Tab. VIII, the larger M is, the better the performance, but the more computation and memory are required to perform inference, so this trade-off has to be balanced.

TABLE VII: IoU performance with varying values of N. **Bold** indicates best performance.

N	Soil IoU	Vegetation IoU	mIoU
12	98.0	77.9	87.9
24	98.7	83.1	90.9
48	98.9	84.4	91.7
56	98.9	84.5	91.7
96	**99.0**	**84.8**	**91.9**

TABLE VIII: IoU performance with varying values of M. **Bold** indicates best performance.

M	Soil IoU	Crop IoU	Weed IoU	mIoU
100	**98.9**	**74.7**	10.4	61.4
250	**98.9**	67.9	11.5	59.4
500	**98.9**	73.5	**12.8**	**61.7**

VI. Conclusion

We presented a novel approach for zero-shot crop-weed semantic segmentation without requiring additional labels. We base our approach on the idea that crop plant features are the most commonly occurring features, while weed features are less common. Our method exploits foundation models SAM [17] and BioCLIP [34] to perform inference without any additional labels or retraining. We implemented and evaluated our approach on a diverse choice of several datasets. Our experiments demonstrate that our unsupervised anomaly detection method is a viable approach for performing crop-weed semantic segmentation without labels.

Our approach leverages the idea that crop plants are the most commonly occurring plants in the field, which leads to a limitation in the case of a weed infestation where specific weed varieties overwhelm the fields. In such cases, our current approach will not be as effective. In our future work, we aim to overcome these limitations by incorporating a small number of weak labels to improve our performance.

References

[1] M. Abdulsalam, U. Zahidi, B. Hurst, S. Pearson, G. Cielniak, and J. Brown. Unsupervised Tomato Split Anomaly Detection Using Hyperspectral Imaging and Variational Autoencoders. In *Proc. of the Europ. Conf. on Computer Vision Workshops*, 2025.

[2] A. Ahmadi, M. Halstead, and C. McCool. Virtual Temporal Samples for Recurrent Neural Networks: Applied to Semantic Segmentation in Agriculture. In *Proc. of the Symp. of the German Association for Pattern Recognition (DAGM)*, 2021.

[3] S. Akcay, A. Atapour-Abarghouei, and T.P. Breckon. GANomaly: Semi-supervised Anomaly Detection via Adversarial Training. In *Proc. of the Asian Conf. on Computer Vision (ACCV)*, 2019.

[4] J. Bendig, K. Yu, H. Aasen, A. Bolten, S. Bennertz, J. Broscheit, M.L. Gnyp, and G. Bareth. Combining UAV-based Plant Height from Crop Surface Models, Visible, and Near Infrared Vegetation Indices for Biomass Monitoring in Barley. *International Journal of Applied Earth Observation and Geoinformation*, 39:79–87, 2015.

[5] C.I. Bercea, B. Wiestler, D. Rueckert, and J. Schnabel. Diffusion Models with Implicit Guidance for Medical Anomaly Detection. In *Proc. of the Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2024.

[6] R. Bertoglio, A. Mazzucchelli, N. Catalano, and M. Matteucci. A Comparative Study of Fourier Transform and CycleGAN as Domain Adaptation Techniques for Weed Segmentation. *Smart Agricultural Technology*, 4:100188, 2023.

[7] K. Buddha, H.J. Nelson, D. Zermas, and N. Papanikolopoulos. Weed Detection and Classification in High Altitude Aerial Images for Robot-Based Precision Agriculture. In *Proc. of the Mediterranean Conf. on Control and Automation (MED)*, 2019.

[8] A. Carraro, M. Sozzi, and F. Marinello. The Segment Anything Model (SAM) for Accelerating the Smart Farming Revolution. *Smart Agricultural Technology*, 6:100367, 2023.

[9] T. Choi, O. Would, A. Salazar-Gomez, X. Liu, and G. Cielniak. Channel Randomisation: Self-supervised Representation Learning for Reliable Visual Anomaly Detection in Speciality Crops. *Computers and Electronics in Agriculture*, 226:109416, 2024.

[10] Y.L. Chong, J. Weyler, P. Lottes, J. Behley, and C. Stachniss. Unsupervised Generation of Labeled Training Images for Crop-Weed Segmentation in New Fields and on Different Robotic Platforms. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2024.

[11] T. Defard, A. Setkov, A. Loesch, and R. Audigier. PaDiM: A Patch Distribution Modeling Framework for Anomaly Detection and Localization. In *Proc. of the Intl. Conf. on Pattern Recognition (ICPR) Workshops and Challenges*, 2021.

[12] M. Fawakherji, A. Youssef, D. Bloisi, A. Pretto, and D. Nardi. Crop and Weeds Classification for Precision Agriculture Using Context-Independent Pixel-Wise Segmentation. In *Proc. of the IEEE Intl. Conf. on Robotic Computing (IRC)*, 2019.

[13] A.A. Gitelson, Y.J. Kaufman, R. Stark, and D. Rundquist. Novel Algorithms for Remote Estimation of Vegetation Fraction. *Remote Sensing of Environment*, 80(1):76–87, 2002.

[14] 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.

[15] X. Jiang, J. Liu, J. Wang, Q. Nie, K. Wu, Y. Liu, C. Wang, and F. Zheng. SoftPatch: Unsupervised Anomaly Detection with Noisy Data. In *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*, 2024.

[16] S. Kawashima and M. Nakatani. An Algorithm for Estimating Chlorophyll Content in Leaves Using a Video Camera. *Annals of Botany*, 81(1):49–54, 1998.

[17] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.Y. Lo, et al. Segment Anything. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2023.

[18] 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.

[19] S. Li, J. Cao, P. Ye, Y. Ding, C. Tu, and T. Chen. ClipSAM: CLIP and SAM collaboration for zero-shot anomaly segmentation. *Neurocomputing*, 618:129122, 2025.

[20] 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(1):20–34, 2020.

[21] M. Louhaichi, M.M. Borman, and D.E. Johnson. Spatially Located Platform and Aerial Photography for Documentation of Grazing Impacts on Wheat. *Geocarto International*, 16(1):65–70, 2001.

[22] F. Magistri, J. Weyler, D. Gogoll, P. Lottes, J. Behley, N. Petrinic, and C. Stachniss. From One Field to Another – Unsupervised Domain Adaptation for Semantic Segmentation in Agricultural Robotics. *Computers and Electronics in Agriculture*, 212:108114, 2023.

[23] W. Mao, Y. Wang, and Y. Wang. Real-Time Detection of Between-Row Weeds Using Machine Vision. In *Proc. of American Society of Agricultural Engineers (ASAE) Annual Meeting*, 2003.

[24] C. McCool, J. Beattie, J. Firn, C. Lehnert, J. Kulk, R. Russell, T. Perez, and O. Bawden. Efficacy of Mechanical Weeding Tools: A Study into Alternative Weed Management Strategies Enabled by Robotics. *IEEE Robotics and Automation Letters (RA-L)*, 3(2):1184–1190, 2018.

[25] G.E. Meyer and J.C. Neto. Verification of Color Vegetation Indices for Automated Crop Imaging Applications. *Computers and Electronics in Agriculture*, 63(2):282–293, 2008.

[26] J.C. Neto. *A combined statistical-soft computing approach for classification and mapping weed species in minimum-tillage systems*. PhD thesis, The University of Nebraska-Lincoln, 2004.

[27] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning Transferable Visual Models from Natural Language Supervision. In *Proc. of the Intl. Conf. on Machine Learning (ICML)*, 2021.

[28] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proc. of the Intl. Conf. on Machine Learning (ICML)*, 2021.

[29] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, and L. Zhang. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. *arXiv preprint*, arXiv:2401.14159, 2024.

[30] E. Romera, J.M. Álvarez, L.M. Bergasa, and R. Arroyo. ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation. *IEEE Trans. on Intelligent Transportation Systems (TITS)*, 19(1):263–272, 2018.

[31] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler. Towards Total Recall in Industrial Anomaly Detection. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[32] C. Schuhmann, R. Beaumont, R. Vencu, C.W. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S.R. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev. LAION-5B: An Open Large-scale Dataset for Training Next Generation Image-Text Models. In *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022.

[33] D. Steininger, A. Trondl, G. Croonen, J. Simon, and V. Widhalm. The CropAndWeed Dataset: A Multi-Modal Learning Approach for Efficient Crop and Weed Manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023.

[34] S. Stevens, J. Wu, M.J. Thompson, E.G. Campolongo, C.H. Song, D.E. Carlyn, L. Dong, W.M. Dahdul, C. Stewart, T. Berger-Wolf, W. lun Chao, and Y. Su. BioCLIP: A Vision Foundation Model for the Tree of Life. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[35] C.J. Tucker. Red and Photographic Infrared Linear Combinations for Monitoring Vegetation. *Remote Sensing of Environment*, 8(2):127–150, 1979.

[36] Y. Wang, Z. Yang, G. Kootstra, and H.A. Khan. The Impact of Variable Illumination on Vegetation Indices and Evaluation of Illumination Correction Methods on Chlorophyll Content Estimation Using UAV Imagery. *Plant Methods*, 19(1):51, 2023.

[37] J. Weyler, T. Läbe, J. Behley, and C. Stachniss. Panoptic Segmentation with Partial Annotations for Agricultural Robots. *IEEE Robotics and Automation Letters (RA-L)*, 9(2):1660–1667, 2024.

[38] J. Weyler, T. Läbe, F. Magistri, J. Behley, and C. Stachniss. Towards Domain Generalization in Crop and Weed Segmentation for Precision Farming Robots. *IEEE Robotics and Automation Letters (RA-L)*, 8(6):3310–3317, 2023.

[39] J. Weyler, F. Magistri, E. Marks, Y.L. Chong, M. Sodano, G. Roggiolani, N. Chebrolu, C. Stachniss, and J. Behley. PhenoBench — A Large Dataset and Benchmarks for Semantic Image Interpretation in the Agricultural Domain. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 1–12, 2024.

[40] D.M. Woebbecke, G.E. Meyer, K. Von Bargen, and D.A. Mortensen. Color Indices for Weed Identification Under Various Soil, Residue, and Lighting Conditions. *Trans. of the American Society of Agricultural Engineers*, 38(1):259–269, 1995.

[41] J. Wyatt, A. Leach, S.M. Schmon, and C.G. Willcocks. AnoDDPM: Anomaly Detection with Denoising Diffusion Probabilistic Models using Simplex Noise. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops*, 2022.

[42] R. Zenkl, R. Timofte, N. Kirchgessner, L. Roth, A. Hund, L. Van Gool, A. Walter, and H. Aasen. Outdoor Plant Segmentation with Deep Learning for High-Throughput Field Phenotyping on a Diverse Wheat Dataset. *Frontiers in Plant Science*, 12:774068, 2022.

[43] Z. Zhang, E. Kayacan, B. Thompson, and G. Chowdhary. High Precision Control and Deep Learning-Based Corn Stand Counting Algorithms for Agricultural Robot. *Autonomous Robots*, 44(7):1289–1302, 2020.

CERTIFICATE OF REPRODUCIBILITY

The authors of this publication declare that:

1) The software related to this publication is distributed in the hope that it will be useful, support open research, and simplify the reproducibility of the results but it comes without any warranty, and without the implied warranties of merchantability and fitness for a particular purpose.

2) *Linn Chong* primarily developed the implementation related to this paper. This was done on Ubuntu 20.04.

3) *Lucas Nunes* verified that the code can be executed on a machine that follows the software specification given in the Git repository available at:

```
https://github.com/PRBonn/WeedsAreWeird
```

4) *Lucas Nunes* verified that the experimental results presented in this publication can be reproduced using the implementation used at submission, which is labeled with a tag in the Git repository and can be retrieved using the command:

```
git checkout master
```