Gradient and Log-based Active Learning for Semantic Segmentation of Crop and Weed for Agricultural Robots

Rasha Sheikh Andres Milioto Philipp Lottes Cyrill Stachniss Maren Bennewitz Thomas Schultz

***Abstract*— Annotated datasets are essential for supervised learning. However, annotating large datasets is a tedious and time-intensive task. This paper addresses active learning in the context of semantic segmentation with the goal of reducing the human labeling effort. Our application is agricultural robotics and we focus on the task of distinguishing between crop and weed plants from image data. A key challenge in this application is the transfer of an existing semantic segmentation CNN to a new field, in which growth stage, weeds, soil, and weather conditions differ. We propose a novel approach that, given a trained model on one field together with rough foreground segmentation, refines the network on a substantially different field providing an effective method of selecting samples to annotate for supporting the transfer. We evaluated our approach on two challenging datasets from the agricultural robotics domain and show that we achieve a higher accuracy with a smaller number of samples compared to random sampling as well as entropy based sampling, which consequently reduces the required human labeling effort.**

I. INTRODUCTION

The ability to interpret the scene in front of a robot is key for intelligent behavior in several applications. For example, precision farming robots need to know which type of plant they perceive or autonomous cars need to know which object in their surroundings is a car, a pedestrian, or a cyclist. These classification or semantic segmentation tasks are typically tackled using convolutional neural networks (CNNs) operating on image data. In order to perform well, neural networks need to be trained with appropriately annotated datasets.

The performance of most supervised learning approaches and especially deep learning systems is related to the quality and quantity of training data. Annotated training data, however, has a high cost as often a larger number of labeled training data is required. In this work, we focus on optimizing the training set generation for semantic segmentation of image data obtained from a mobile robot. Semantic segmentation refers to the task of computing a pixel-wise labeling of the images. More concretely, we address the agricultural robotics application in which robots should perform automated weed control. For the semantic segmentation, this means that we need to compute the semantic label "crop", "weed", or "misc" for each pixel in the image. This task is particularly challenging as the field conditions often change substantially between years, regions, weather, and soil conditions as can be seen in Figure 1.

One solution to adapt and refine existing semantic segmentation systems to new field conditions is through additional labeled data from the new field. As these new annotations need to be executed at the end-users site, one is interested in keeping this effort as low as possible. Given annotated data on one agricultural field and a CNN that was trained on it, we address the problem of transferring this knowledge to new fields with minimum effort. Datasets from different fields reveal different crop and weed statistics. They often differ by soil type, weather condition, or various small objects that can be found on the ground, such as stones, dried vegetation, or marks from agricultural machines, i.e., patterns that are neither crop nor weed. Additionally, the robot can acquire images of plants at a certain growth stage in one field, while the growth state on the target field is different. Lastly, artifacts such as contrast changes can be found in the camera images captured from the various locations. As illustrated by Lottes *et al.* [19], [20], these conditions make it difficult to simply reuse a previously trained network from one field and infer the labels on another.

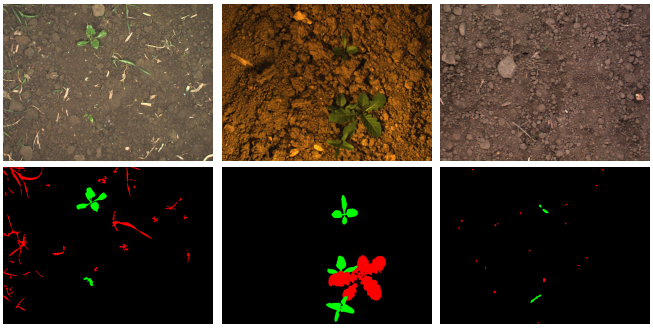

Fig. 1. Sample images from the Bonn, Stuttgart, and Zurich sugar beet datasets in the first, second, and third column, respectively. The first row shows the RGB images and the second row shows their annotations (green denotes crop while red denotes weed). As can be seen, the appearance differs substantially.

The contribution of this work is to introduce and compare three active learning strategies that intelligently pick images taken under new conditions to re-train an existing network: The first one picks samples based on a log-space ranking of their loss with respect to pseudo labels. The second and third approaches select training samples that are expected to have a maximum effect on the network weights. Even though similar ideas have been explored for active learning in other application contexts, it is non trivial to apply them for semantic segmentation. An important technical novelty in our work is to exploit a pseudo ground truth, which we obtain with very weakly supervised segmentation. Our approach selects samples in batches, each time refining the network, then computing a new ranking of the unlabeled data. The best samples are then selected and the network is re-trained. To compute the real gradients, corresponding ground truth

All authors are with the University of Bonn, Germany. This work has partly been supported by the German Research Foundation under Germany's Excellence Strategy, EXC-2070 - 390732324 (PhenoRob).

data is needed. Thus, in our approach, we approximate the ground truth as the result of unsupervised segmentation to estimate the gradient. We evaluated our framework using three distinctive sugar beet datasets [5] that have different characteristics. Our results indicate that our method produces a higher accuracy on the datasets with a fewer number of samples compared to random sampling for annotation as well as entropy based sampling.

II. RELATED WORK

Several works focusing on the elimination or reduction of herbicide use, through the incorporation of autonomous ground robots in crop fields, have been introduced to the community in the last years [7], [16], [21]. A key component of each of these unmanned platforms is a core perception system that has the ability to accurately distinguish crops from weeds in order to effectively and selectively apply the desired individual treatment [18], [22], [23], [24], [27]. These systems allow autonomous robots to perform actuation in the fields without human supervision, treating each plant individually. All of the works referenced, however, are based on supervised learning approaches which take large amounts of pixel-accurate hand-labeled images for training. Accordingly, one of the main bottlenecks of these visual processing pipelines is the amount of expensive labeled training data required to deploy them in real agricultural fields, which often limits their applicability. In order to tackle this data starvation problem, we propose an active learning based solution.

Numerous works on general active learning have been presented in the community [30], [11], [12], [36]. The most common measures for selecting samples are based on the uncertainty of the network [38], [35], [10], [33] and diversity [38], [8], [14]. Sener *et al.* [29] assert based on the experiments they performed that uncertainty based approaches are not effective for active learning with CNNs. They hypothesize that this is not due to the inaccurate estimate of uncertainty by the network, rather to the ineffectiveness of uncertainty based approaches to cover the space of image features. The Expected Model Output Change Principle (EMOC) developed by Freytag *et al.* [9] tries to avoid selecting samples that are redundant and Käding *et al.* [14] follow this approach with deep neural networks. This principle measures how a model would perform with and without the candidate sample. Given that the labels are unknown, a marginalization over the possible labels is needed. Uncertainty estimation for active learning can be performed using Monte-Carlo dropout as in [10] or with an ensemble of deep networks. Beluch *et al.* [3] compare both of these approaches on different datasets. They found that an ensemble of deep classifiers has a superior performance even with a smaller number of models. They conclude that Monte-Carlo dropout approaches suffer from a lower diversity and a smaller model capacity.

Weakly supervised segmentation is an active research topic [34], [1], [32], [15]. In the context of self-learning, Zhang *et al.* [37] use labels obtained with K-means graph cuts as ground truth for their network. The predictions produced by the model are then used as the target labels for the next iteration of the process.

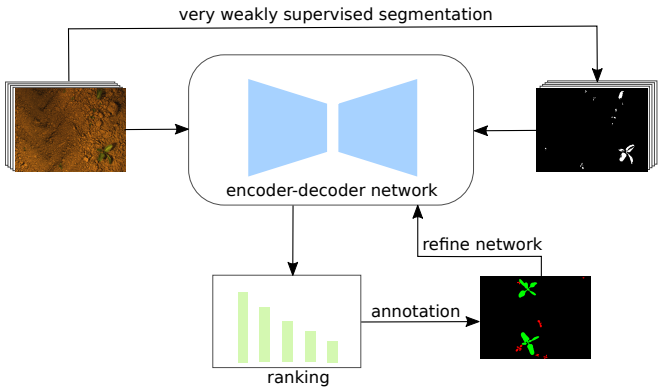

Fig. 2. Overview of our approach. The key idea is that we first perform a very weakly supervised segmentation to obtain pseudo ground truth. Given the labels and different ranking measures obtained from the network, we rank the unlabeled samples and pick them accordingly for annotation. Those samples are then used to refine the entire network.

The works mentioned previously and the current state-of-the-art methods for active learning including [10], [3], [28], [36] are either more suitable for tasks other than pixel-wise semantic segmentation of images with CNNs and/or are memory and computationally expensive. Differently, we experiment with approaches that directly measure how annotated samples can affect the gradients. We use labels obtained with very weak supervision as pseudo ground truth and compute the gradients w.r.t the weights. We then refine a pre-trained network with the newly annotated samples in an iterative manner. Our intuition for using gradients is driven by the observation that the greater the mismatch is between the predicted segmentation and the ground truth, the larger the change is to the weights. This is in contrast to most of the approaches mentioned earlier that rely on the confidence of the network which may not be the best indication of the best samples to choose for annotation, as the network output might actually be correct although the network is uncertain about it.

Previous work, such as the Expected Gradient Length (EGL) [13], [31], has explored how changes in model parameters can be exploited for sample selection. However, it computes the expectation of the gradient norm over all possible annotations, which would be prohibitively expensive for pixel-wise semantic segmentation of images. We instead compute gradients from rough foreground/background segmentation. Du *et al.* [6] use gradient similarity to determine when an auxiliary task is helpful for transfer learning to the main task and when it can be hurtful. Although in our work, the weakly supervised setting can be seen as an auxiliary task, we only use the gradients computed there as a guidance to choose samples for annotations. These gradients are not used to measure similarity with those of the main task nor are the parameters of the main task updated with those gradients.

III. OUR APPROACH TO EFFECTIVE SAMPLE SELECTION

Figure 2 shows an overview of our framework. The key idea of our approach is to perform a very weakly supervised

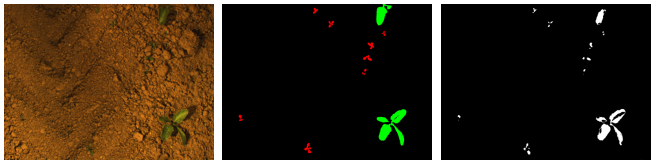

Fig. 3. Very weakly supervised segmentation used as pseudo ground truth by our approach. Left: Input image. Middle: Ground truth semantic segmentation; Right: Foreground segmentation of vegetation provided by k-means clustering. Note that only such a rough segmentation as pseudo ground truth is enough for our approach.

segmentation to obtain pseudo ground truth. Given the labels and different measures produced by the network, we rank the unlabeled samples and pick them accordingly for annotation. These are then used to refine the entire network.

Our CNN for semantic segmentation relies on Bonnet [25]. The used network is based on SegNet [2] and ENet [26]. It has an encoder-decoder structure with a total of 25 [5x5] convolutional layers. It uses batch normalization, residual connections, ReLU as the non-linearity layer, and the focal loss function [17]. As input to our network, we only use the standard RGB channels of a camera.

In order to perform the semantic segmentation in sugar beet field for agricultural robotics tasks, we train our model on the Bonn sugar beet dataset [5]. We then refine the trained model on other datasets by incrementally selecting batches of samples. The datasets differ in their crop/weed statistics and the images acquired with the cameras also differ in their illumination. Therefore, simply running the trained model to segment the vegetation in other fields does not work.

We compare three different approaches to sample selection for active learning. Our main technical contribution is the generation of a pseudo ground truth (Sec. III-B) and its use for loss-based (Sec. III-C), as well as two gradient-based approaches for sample selection (Sec. III-D and Sec. III-E).

A. Setup

We evaluate our different approaches by first training a network on the Bonn sugar beet dataset then refining it on the Stuttgart and Zurich datasets separately. To refine the network we pick unlabeled samples in batches of 10 using one of the methods described in this section. Once the samples are annotated, they are given to the network. We repeat this process iteratively, each time refining the network on all of the newly annotated samples.

B. Generation and Use of Pseudo Ground Truth

Our three main methods make use of "pseudo ground truth" foreground-background segmentation masks, which we obtain by clustering the values of the RGB channels. An example is shown in Figure 3. We run k-means to determine 20 cluster representatives from 10 randomly selected images. After viewing a single image that contains all 20 clusters, a human annotator chooses which clusters represent vegetation. In our experiments, it was enough to select two clusters. Therefore, the human annotation effort that is required to obtain the pseudo ground truth amounts to a few seconds for a complete new dataset. In accordance with previously used terminology [37], we refer to this as very weak supervision. Figure 3 shows an image, its ground truth and the foreground segmentation (pseudo ground truth) provided by clustering. It is an important finding from our experiments that a rough and easy to compute segmentation is sufficient for the purpose of selecting images for annotation. This makes our proposed gradient-based approach feasible in practice.

In order to compute a loss from the network output, which includes three classes, and the pseudo ground truth, which merely includes two, one might combine crop and weed into a single foreground class, or treat the foreground class as a specific type of vegetation (i.e., crop or weed). We tried all three options and found that treating the foreground from the pseudo ground truth as crop empirically produced the best result. We emphasize that the pseudo ground truth is only used to select training samples that should be annotated; the network weights are updated based on manual annotations of the selected samples, which include all three classes.

In our agricultural application, the number of true classes (3) is not much higher than the number of classes (2) in our pseudo-ground truth. Naturally, in a different semantic segmentation task, the number of classes could be higher and might require generating a pseudo-ground truth with a larger number of classes. Our method here uses a simple clustering mechanism but other unsupervised or weakly supervised methods can also be used to generate pseudo-labels with a higher number of classes that can be later used to compute the gradients for sample selection.

C. Sample Selection Using Loss

The loss of the network is an indication of the segmentation error. Given that training neural networks with backpropagation is driven by the loss, it also provides a useful cue as to which samples the network will most benefit from. We compute the focal loss [17] based on the pseudo ground truth.

We found that training only on the images with the highest loss values did not generalize well. This could indicate that they are not representative enough of the overall dataset. Therefore, we instead employ a scheme that samples images with a diverse range of loss values, but prefers those with higher losses. To this end, we sort the images by their loss in a descending order, and then select them uniformly on a logarithmic scale. Specifically, we compute index i of the n-th sample as:

$$i = \lfloor |P|^{n/(|S|-1)} \rfloor - 1, \qquad n \in \{0, 1, \ldots, |S|-1\} \quad (1)$$

where $|S|$ is the number of samples to be selected and $|P|$ is the size of the images pool. Since the samples are sorted, this approach would more heavily select those that have higher loss values while not completely discarding images that the network is performing well on.

D. Sample Selection Using Norm of Gradients

For this approach and the following one, we pick those samples for annotation that might have the largest impact on the network weights. The norm of the network gradients is a

measure that is indicative of which samples will affect the weights more than others. Although the loss and norm of gradients are correlated, there are instances where the loss could be high for certain samples, yet the gradient is locally small. This depends on the loss function and the state of the current network parameters.

As in the previous approach, we use labels from very weakly supervised segmentation as pseudo ground truth. We run the network on the training images for one epoch (to maintain computational efficiency) and compute the gradients. Again we note that this step is only used to compute the gradients but the network weights remain unchanged. Once we have the gradients, we compute the L_2 norm of those in the last two layers of the network (the classifier layer and the one immediately before it):

$$n_g(\mathbf{x}) = \left\| \nabla_{w_f} \mathcal{L}(\mathbf{x}) \right\|, \tag{2}$$

where $\mathbf{x}$ is the image and w are the weights of the final two layers. The images are sorted based on this measure in a descending order and again we pick samples on a log-space scale afterwards as explained earlier.

E. Sample Selection Using Gradient Projection

The log-space in the previous approaches was used to ensure there is enough diversity among the samples so that the network does not overfit on them and can generalize to unseen data. Here we use a different method that relies on the space spanned by the gradients where we project onto the orthogonal complement of the gradients of the selected samples. For every picked sample, we project the gradients of all remaining samples onto the selected sample gradient. We then subtract the projected gradient from the original gradients. The residual we are left with indicates which samples have the strongest remaining effect on the weights after accounting for the already selected samples. This can be formulated as:

$$n_p(\mathbf{x}) = \left\| \mathbf{g_x} - \sum_{i=1}^{S} \frac{\langle \mathbf{g_i}, \mathbf{g_x} \rangle}{\langle \mathbf{g_i}, \mathbf{g_i} \rangle} \mathbf{g_i} \right\|, \tag{3}$$

where $\mathbf{x}$ is the image, $\mathbf{g_i}$ is the gradient of the ith sample out of S previously selected samples, and $\mathbf{g_x}$ is the gradient of the current sample. We select samples one by one, each time sorting them according to this measure and choosing the one with the highest norm of the residual. To pick the first sample, we choose that with the highest norm of the gradient.

IV. Experimental Evaluation

In this section, we demonstrate the effectiveness of the approaches we designed for active learning and evaluate the performance of the different sample selection methods on different datasets, and compare them to random and entropy based approaches.

TABLE I

Datasets Statistics of Crop and Weed Plants

	Bonn	Stuttgart	Zurich
Images	8230	2584	2577
Crop pixels	2.0%	1.5%	0.4%
Weed pixels	0.3%	0.7%	0.1%

TABLE II

IoU without any refinement (lower bound) and IoU when training on the whole dataset (upper bound).

	No Refinement	Fully supervised
Stuttgart	0.3429	0.7989
Zurich	0.3595	0.7024

A. Datasets

The datasets we used were acquired with a Bosch Deepfield Robotics UGV. The robot was developed to assist in several agricultural applications, including mechanical weed control and selective herbicide spraying [5]. It is equipped with multiple sensors such as cameras, GPS trackers, and 3D laser sensors. For our experiments we use the RGB data provided by the JAI AD-130GE camera.

The data was captured in three different fields: Bonn and Stuttgart in Germany, and Zurich in Switzerland. The datasets have weed and crop plants at different stages of growth. Figure 1 shows sample images from the different datasets. The images vary in their illumination, soil type, and class statistics, hence the need for transfer learning. The images have been annotated into three classes: weed, crop, and soil/misc. Table I shows the number of images in each dataset and the ratio of foreground pixels. It can be clearly seen that there is a high imbalance of classes in the data. We follow the approach of [24] and split the new dataset into three sets: 40% for training, 10% for validation, and 50% for testing. The samples are picked from the training set. All experiments were conducted on four Nvidia Titan X GPUs.

B. Re-Training Performance

The experiments in this section are designed to show how the proposed sample selection strategies impact the performance of the network in the new environment. For quantifying the performance, we use the mean Intersection over Union (mIoU) as the performance measure. To provide the lower and upper bounds for the methods, we list in Table II the mIoU for each dataset when running the model without any refinement as well as when training on all of the samples.

Figures 4 and 5 show the performance on the Stuttgart and Zurich datasets when selecting samples for annotation with different methods. As baselines we include random sampling, and selecting samples that have the highest entropy ([4], [38]):

$$H(\mathbf{x}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c} p(c \mid x_i) \log p(c \mid x_i), \tag{4}$$

where x_i is pixel i in image $\mathbf{x}$, c is the class and N is the number of pixels in the image.

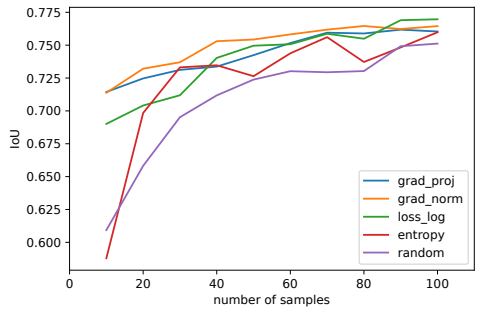

Fig. 4. Pixel-wise mean IoU on the Stuttgart dataset. Running the model without any new annotations yields an IoU of 0.34. Running the model on the whole dataset yields an IoU of 0.79. Gradient-based approaches can reach 90% of the fully supervised performance with 10 samples.

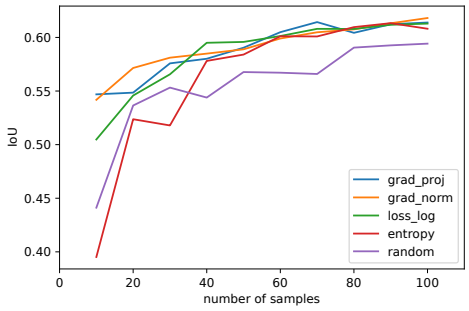

Fig. 5. Pixel-wise mean IoU on the Zurich dataset. Running the model without any new annotations yields an IoU of 0.36. Running the model on the whole dataset yields an IoU of 0.70. Gradient-based approaches can reach 77% of the fully supervised performance with 10 samples.

A few observations can be made from the figures: the effect of the sampling method is stronger when only a few images are selected. As the model is trained on more and more samples, the accuracy plateaus as expected and the variation between the different methods decreases. It can be noted however that random sampling has a lower performance even with a greater number of images.

The overall performance on the Stuttgart dataset is better than that on the Zurich dataset. This can be attributed to the different class statistics of the two datasets. As can be seen in Table I, the Stuttgart dataset has a larger percentage of crop and weed pixels compared to the Zurich dataset. This allows the model to better distinguish between the different classes. This observation is also supported by the fully supervised performance shown in Table II where a higher IoU can be obtained on the Stuttgart dataset.

When training the model with only a handful of images, 10 or 20 images, the methods that take into account the impact of the samples on the weights lead to better generalization to the rest of the unseen data. In particular, ranking samples by projecting out gradients results in higher mIoU on both datasets. With 10 samples, which would amount to roughly 1% of the training dataset size, we can achieve 90% of the fully supervised performance (Table II) on the Stuttgart dataset, compared to 76% with random selection. On the Zurich dataset, we can achieve 77% of the fully supervised performance compared to 63% with random selection.

TABLE III

OBJECT-WISE PERFORMANCE ON THE STUTTGART AND ZURICH DATASETS RESPECTIVELY. EACH ROW SHOWS THE PERFORMANCE AFTER SELECTING 10 SAMPLES WITH THE DIFFERENT METHODS AND REFINING THE NETWORK. RUNNING THE MODEL WITHOUT ANY NEW ANNOTATIONS YIELDS AN ACCURACY OF 0.15 ON STUTTGART AND 0.33 ON ZURICH.

Samples No.	Random	Entropy	Loss	Gradient Norm	Gradient Proj.
10	0.6920	0.6890	0.7882	0.8040	**0.8196**
20	0.7402	0.8050	0.7769	0.8350	**0.8404**
30	0.8138	0.8300	0.7950	0.8461	**0.8470**
40	0.8254	0.8463	0.8555	**0.8682**	0.8252
50	0.8225	0.8405	0.8523	**0.8599**	0.8278

Samples No.	Random	Entropy	Loss	Gradient Norm	Gradient Proj.
10	0.7552	0.7879	0.7697	**0.8354**	0.8025
20	0.7971	0.8212	0.8189	**0.8768**	0.8170
30	**0.8591**	0.7884	0.8321	0.8553	0.8299
40	0.8575	**0.8711**	0.8610	**0.8711**	0.8479
50	0.8593	0.8688	0.8636	**0.8852**	0.8784

To further quantify the performance of our approach, we use the object-wise metric defined by Milioto *et al.* [24], where the accuracy is measured for objects larger than 50 pixels. Since the target application is weeding with agricultural robotics, this metric is more directly useful than pixel-wise performance. Table III shows how our approach performs on the Stuttgart and Zurich datasets. Each row shows the mean accuracy when selecting n samples with different methods. For comparison, random and entropy based sampling are shown in the first and second columns respectively.

C. Comparison to Other Baselines

To gain more insight into what our baselines are, we ran additional experiments with the results shown in Table IV.

In the first row, we ran an experiment where we trained the model with the pseudo ground truth first and picked samples randomly afterwards. We found that it performs slightly better than when picking random samples directly (0.64 vs. 0.61) but still worse than our log and gradient based methods (e.g. 0.64 vs. 0.71 for the gradient-norm approach). Although pre-training with the pseudo ground truth allows the network to distinguish foreground vegetation from background, the task at hand is to learn three classes and more importantly distinguish crop from weed. Therefore for all experiments, we refine the model without pre-training on the foreground masks.

In the second row, we run an "oracle" experiment. We compute the difference between the parameters of the model without any refinement and the parameters of the fully supervised model. We then find samples with gradients that align with the parameters difference. This experiment is not intended for sample selection, rather to know if the framework had complete knowledge of how the gradients should look

TABLE IV

ADDITIONAL BASELINES FOR TRAINING WITH 10 SAMPLES ON THE STUTTGART DATASET. COMPARE WITH FIG. 4.

Method	mIoU
Random-pseudo ground truth	0.6448
Align with parameters difference (oracle)	0.7010

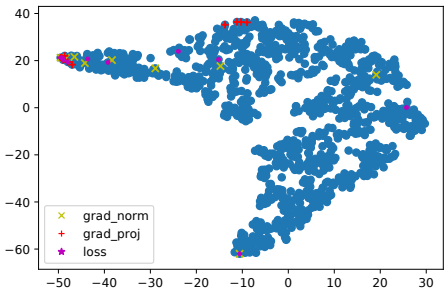

Fig. 6. t-SNE of the images gradients on the Stuttgart dataset. Each point represents the 2-D embedding of the gradient vector. The first 10 samples selected by each method are shown in different colors.

like, would it be able to pick better samples. We found that the oracle performance is similar to our gradient-based approaches after seeing 10 new samples. This implies that the gradient-based approaches are bounded by this performance. Substantially improving upon their performance might require exploiting additional knowledge from the model, possibly with the aid of unsupervised segmentation.

D. Inspecting t-SNE of Samples Gradients

To further analyze the ranking methods and inspect potential patterns in the different sampling approaches, we plot the t-distributed Stochastic Neighbor Embedding (t-SNE) of the gradients in Figure 6. Each circle denotes the 2-D embedding of the gradient of a single image before picking the first 10 samples. Samples selected by each method are shown in different colors. As explained in Section III-D and III-E, we combined the idea of gradient-based selection with two alternative approaches to achieving diversity in the selected images: picking on a log scale, or projecting out gradients that have been selected previously. In our experiments, both strategies performed well (see Fig. 4 and Fig. 5). When inspecting the gradients of the samples selected, we found that the strongest gradients cluster together, near the top left. Additionally, the gradient projection method selects many points at the boundary of the distribution, suggesting that it might be improved further by adding a mechanism to ensure that selected images are representative of a larger subset in the overall dataset.

E. Performance on Weed and Crop Classes

A more detailed breakdown of the methods performance is shown in Table V. The first table shows the pixel-wise precision and recall on the Stuttgart dataset after selecting the first 10 samples. Both methods, Gradient Norm and Gradient Projection have a high recall and precision of the crop class without degrading those of the weed class. The object-wise performance in the second table further illustrates the effectiveness of these methods. Gradient Norm and Gradient Projection produce high precision and recall for both classes. We observed the same behavior on the Zurich dataset (not included here).

TABLE V

PRECISION AND RECALL ON THE STUTTGART DATASET AFTER SELECTING THE FIRST 10 SAMPLES. THE FIRST TABLE SHOWS THE PIXEL-WISE PERFORMANCE AND THE SECOND TABLE SHOWS THE OBJECT-WISE PERFORMANCE. THE HIGHEST VALUES ALONG A COLUMN ARE IN BOLD AND THE LOWEST IN ITALICS.

	Precision		Recall	
	Weed	Crop	Weed	Crop
Random	*0.4095*	*0.7278*	0.4851	0.6946
Entropy	0.4158	0.7334	*0.4786*	*0.5894*
Loss Log	0.5331	0.8025	0.6179	0.8112
Gradient Norm	**0.5970**	0.8259	0.6136	**0.8402**
Gradient Projection	0.5745	**0.8365**	**0.6564**	0.8212

	Precision		Recall	
	Weed	Crop	Weed	Crop
Random	*0.8723*	0.5740	0.6587	*0.6474*
Entropy	0.8851	*0.5238*	*0.6122*	0.7399
Loss Log	0.9005	0.6898	0.7811	0.7351
Gradient Norm	**0.9090**	**0.7390**	0.7970	**0.7536**
Gradient Projection	0.9030	0.7308	**0.8289**	0.7375

V. CONCLUSION

In this paper, we introduced and compared several active learning approaches that support the adaptation of semantic segmentation networks to new environments. Our approaches effectively select samples from the new environment for user annotation with the goal of maximizing the benefit from a small number of annotated examples. We applied sample selection strategies to the task of crop/weed classification for agricultural robots, as the appearance between agricultural fields often changes substantially such that re-training is needed. We compute pseudo ground truth labels using very weakly supervised segmentation and use those labels to estimate how new, unlabeled samples will affect the weights of the CNN if selected for training. We select the training samples for user annotation based on the estimated effect on the weights and use them to refine the network.

We evaluated the performance gain of our gradient-based and log-based approaches on two agricultural datasets for weed detection. The datasets reveal different characteristics from the dataset on which the network was pretrained. Our results show the effectiveness of our method as it produces higher semantic segmentation accuracies with a small number of training samples, compared to random sampling as well as entropy based sampling. As a result of that, the effort in human annotation is reduced without compromising performance.

REFERENCES

[1] David Acuna, Huan Ling, Amlan Kar, and Sanja Fidler. Efficient interactive annotation of segmentation datasets with polygon-rnn++. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 859–868, 2018.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. on Pattern Analalysis and Machine Intelligence (TPAMI)*, 39(12):2481–2495, 2017.

[3] William H. Beluch, Tim Genewein, Andreas Nürnberger, and Jan M. Köhler. The power of ensembles for active learning in image classification. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 9368–9377, 2018.

[4] Shayok Chakraborty, Vineeth Balasubramanian, Qian Sun, Sethuraman Panchanathan, and Jieping Ye. Active batch selection via convex relaxations with guaranteed solution bounds. *IEEE Trans. on Pattern Analalysis and Machine Intelligence (TPAMI)*, 37(10):1945–1958, 2015.

[5] Nived Chebrolu, Philipp Lottes, Alexander Schaefer, Wera Winterhalter, Wolfram Burgard, and Cyrill Stachniss. Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields. *The Intl. Journal of Robotics Research*, 36(10):1045–1052, 2017.

[6] Yunshu Du, Wojciech M Czarnecki, Siddhant M Jayakumar, Razvan Pascanu, and Balaji Lakshminarayanan. Adapting auxiliary losses using gradient similarity. *arXiv preprint arXiv:1812.02224*, 2018.

[7] Tom Duckett, Simon Pearson, Simon Blackmore, and Bruce Grieve. Agricultural robotics: The future of robotic agriculture. *arXiv preprint*, abs/1806.06762, 2018.

[8] Suyog Dutt Jain and Kristen Grauman. Active image segmentation propagation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2864–2873, 2016.

[9] Alexander Freytag, Erik Rodner, and Joachim Denzler. Selecting influential examples: Active learning with expected model output changes. In *European Conf. on Computer Vision*, pages 562–577, 2014.

[10] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *Proc. of thhe Intl. Conf. on Machine Learning*, pages 1183–1192, 2017.

[11] Isabelle Guyon, Gavin Cawley, Gideon Dror, and Vincent Lemaire. Results of the active learning challenge. In *Proc. of the AISTATS Active Learning and Experimental Design Workshop*, pages 19–45, 2011.

[12] Alex Holub, Pietro Perona, and Michael C. Burl. Entropy-based active learning for object recognition. In *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2008.

[13] Jiaji Huang, Rewon Child, Vinay Rao, Hairong Liu, Sanjeev Satheesh, and Adam Coates. Active learning for speech recognition: the power of gradients. *arXiv preprint arXiv:1612.03226*, 2016.

[14] Christoph Käding, Erik Rodner, Alexander Freytag, and Joachim Denzler. Active and continuous exploration with deep neural networks and expected model output changes. *arXiv preprint arXiv:1612.06129*, 2016.

[15] Suha Kwak, Seunghoon Hong, and Bohyung Han. Weakly supervised semantic segmentation using superpixel pooling network. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[16] Frank Liebisch, Pfeifer Johannes, Raghav Khanna, Philipp Lottes, Cyrill Stachniss, Tillmann Falck, Slawomir Sander, Roland Siegwart, Achim Walter, and Enric Galceran. Flourish – A robotic approach for automation in crop management. In *In Proc. of the Workshop für Computer-Bildanalyse und unbemannte autonom fliegende Systeme in der Landwirtschaft*, 2016.

[17] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proc. of the IEEE Intl. Conf. on Computer Vision (ICCV)*, pages 2999–3007, 2017.

[18] Philipp Lottes, Jens Behley, Nived Chebrolu, Andres Milioto, and Cyrill Stachniss. Joint Stem Detection and Crop-Weed Classification for Plant-specific Treatment in Precision Farming. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2018.

[19] Philipp Lottes, Jens Behley, Andres Milioto, and Cyrill Stachniss. Fully convolutional networks with sequential information for robust crop and

[20] Philipp Lottes and Cyrill Stachniss. Semi-supervised online visual crop and weed classification in precision farming exploiting plant arrangement. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2017.

weed detection in precision farming. *IEEE Robotics and Automation Letters (RA-L)*, 3:3097–3104, 2018.

[21] Chris McCool, James Beattie, Jennifer Firn, Chris Lehnert, Jason Kulk, Raymond Russell, Tristan Perez, and Owen Bawden. Efficacy of mechanical weeding tools: A study into alternative weed management strategies enabled by robotics. *IEEE Robotics and Automation Letters (RA-L)*, 2018.

[22] Chris McCool, Tristan Perez, and Ben Upcroft. Mixtures of Lightweight Deep Convolutional Neural Networks: Applied to Agricultural Robotics. *IEEE Robotics and Automation Letters (RA-L)*, 2017.

[23] Andres Milioto, Philipp Lottes, and Cyrill Stachniss. Real-time blob-wise sugar beets vs weeds classification for monitoring fields using convolutional neural networks. In *Proc. of the Intl. Conf. on Unmanned Aerial Vehicles in Geomatics*, 2017.

[24] Andres Milioto, Philipp Lottes, and Cyrill Stachniss. Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in cnns. In *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 2229–2235, 2018.

[25] Andres Milioto and Cyrill Stachniss. Bonnet: An open-source training and deployment framework for semantic segmentation in robotics using cnns. *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2019.

[26] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*, 2016.

[27] Inkyu Sa, Marija Popovic, Raghav Khanna, Zetao Chen, Philipp Lottes, Frank Liebisch, Juan Nieto, Cyrill Stachniss, Achim Walter, and Roland Siegwart. WeedMap: A Large-Scale Semantic Weed Mapping Framework Using Aerial Multispectral Imaging and Deep Neural Network for Precision Farming. *Remote Sensing*, 10, 2018.

[28] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

[29] Ozan Sener and Silvio Savarese. A geometric approach to active learning for convolutional neural networks. *arXiv preprint arXiv*, 1708:1, 2017.

[30] Burr Settles. Active learning literature survey. Technical report, Univ. of Wisconsin-Madison, Dep. of Computer Sciences, 2009.

[31] Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. In *Advances in Neural Information Processing Systems*, pages 1289–1296, 2008.

[32] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1818–1827, 2018.

[33] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2017.

[34] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 7268–7277, 2018.

[35] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z. Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *Proc. of the Intl. Conf. on Medical Image Computing and Computer-Assisted Intervention*, pages 399–407, 2017.

[36] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 93–102, 2019.

[37] Ling Zhang, Vissagan Gopalakrishnan, Le Lu, Ronald M Summers, Joel Moss, and Jianhua Yao. Self-learning to detect and segment cysts in lung ct images without manual annotation. In *IEEE Intl. Symposium on Biomedical Imaging (ISBI 2018)*, pages 1100–1103, 2018.

[38] Zongwei Zhou, Jae Shin, Lei Zhang, Suryakanth Gurudu, Michael Gotway, and Jianming Liang. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 7340–7351, 2017.