Vision, Modeling, and Visualization (2019), pp. 1–8
H.-J. Schulz, M. Teschner, and M. Wimmer (Eds.)

# Gradient-based Sample Selection for Transfer Learning of Semantic Segmentation: Application in Agricultural Robotics

Submission ID†

paper1038

**Abstract**

*Annotated datasets are essential for supervised learning. However, annotating large datasets to train deep neural networks that perform well 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 target 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. We propose a novel approach that, given a trained model on one field, refines the network on a substantially different field providing an effective method of selecting samples to annotate for supporting the transfer. Our method takes into account the influence of the so far unlabeled samples on the weights of the network and ranks and selects them accordingly for annotation. We evaluated our approach on two challenging datasets and show that we achieve a higher accuracy with a smaller number of samples compared to randomly selecting samples for annotation as well as uncertainty-based approaches to select images for annotation. Thus, our approach reduces the required human labeling effort.*

**CCS Concepts**

• ***Computing methodologies*** → *Image segmentation;* • ***Applied computing*** → *Agriculture;*

## 1. Introduction

Image segmentation is a classical task in computer vision with many important applications, including environment perception of autonomous systems and medical image analysis. Over the last few years, deep convolutional neural networks (CNNs) have firmly established themselves as the state-of-the-art approach for image segmentation. However, their accuracy crucially depends on the quality and quantity of the available training data.

Having human experts annotate a large number of training images incurs a high cost. In many cases, it is now a limiting factor for achieving high-quality segmentations. Therefore, it has become popular to use transfer learning: Instead of training a neural network from a random initialization, an architecture that has been trained for a different task is re-used, and fine-tuned with a limited number of training samples for the new task. It has been observed that especially the features in early layers tend to be application independent, and that transfer learning has a beneficial effect even when it involves two distant tasks [YCBL14].

Often, many more images are available for the target task than the human has time to annotate. In this scenario, it can be beneficial to rank the unlabeled images with respect to their expected benefit for fine-tuning the network. This is referred to as active learning, since the learning technique actively asks for labels on specific inputs. Most available active learning methods focus on samples that the network is uncertain about. In this work, we propose an alternative strategy that is based on two steps as illustrated in Figure 1: In the first step, with minimal human intervention, a very weakly supervised segmentation of all unlabeled images is achieved. This leads to imperfect results, which will be used as a pseudo ground truth. In the second step, images are selected for annotation if the current network output deviates from their pseudo ground truth. In particular, we propose to use images for which the gradient of the loss with respect to the neural network weights has a large norm, since they are expected to have the largest impact during training. We combine this idea with two different strategies to achieve diversity in the chosen samples. The final refinement is done with the correct (i.e., manual) labels of the selected images.

We evaluate our novel strategy on a specific task from agricultural robotics, which is a perfect domain for transfer learning and active sample selection. In precision farming, robots take an image of a field and need to compute the semantic labels “crop”, “weed”, or “misc” for every pixel, in order to perform automated weed control. A challenge lies in the fact that visual appearance varies substantially between years, regions, weather, and soil conditions. Thus, one often tries to adapt and refine existing semantic segmentation systems to new field conditions through annotated data from the new field.

† Acknowledgement

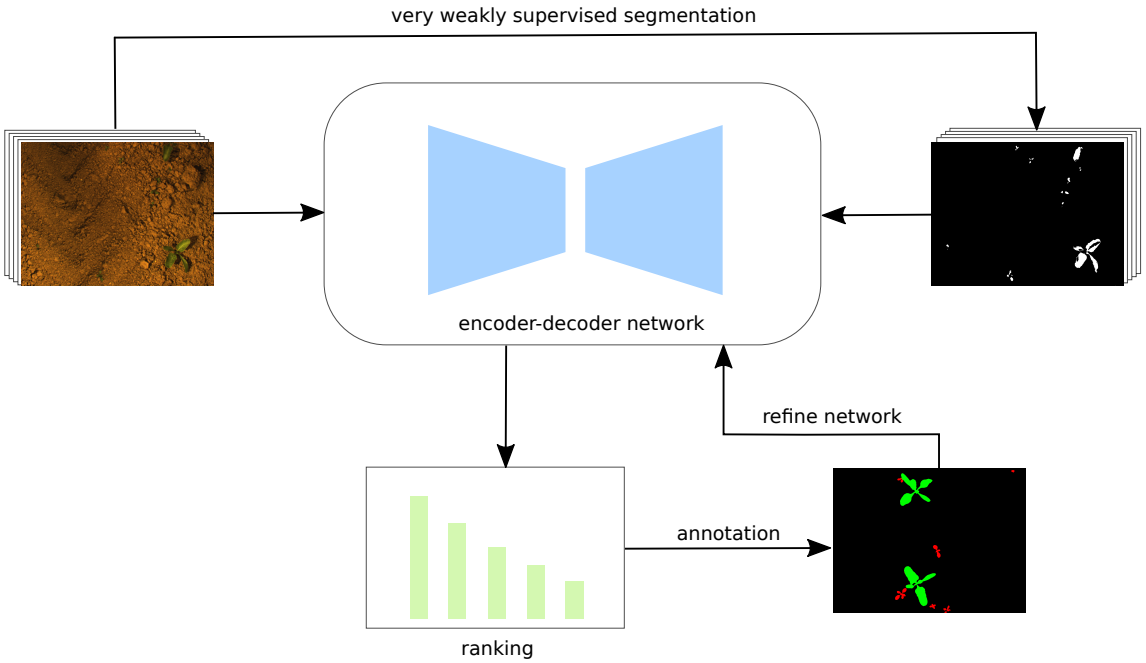



Figure 1: Overview of our system. We first perform very weakly supervised segmentation to obtain pseudo ground truth. Given the labels and different measures produced by the network, we rank the unlabeled samples and pick them accordingly for annotation. These are then used to refine the entire network.

Our results indicate that, when only a small number of new training samples can be annotated, selecting them with our novel gradient-based approach results in a higher accuracy of the final refined CNN when compared to either a random or an uncertainty-based selection of training samples. We also demonstrate that the gradient-approach works better than making the selection simply based on the loss with respect to the same pseudo ground truth.

## 2. Related Work

Numerous works on general active learning have been presented in the community [Set09, GCDL11, HPB08]. The most common measures for selecting samples are based on the uncertainty of the network [ZSZ*17, YZC*17, GIG17, WZL*17] and diversity [ZSZ*17, DJG16, KRFD16]. Sener *et al.* [SS17] 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 by the ineffectiveness of uncertainty based approaches to cover the space of image features.

Weakly supervised segmentation is an active research topic [WXS*18, ALKF18, TDP*18, KHH17]. In the context of self-learning, Zhang *et al.* [ZGL*18] 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.

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 [DPBG18, LJK*16, MBF*18]. 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 [LBC*18, MPU17, MLS17, MLS18, SPK*18]. 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.

Different to the previously mentioned works, 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. Du *et al.* [DCJ*18] 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.

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.

## 3. Effective Sample Selection

Figure 1 shows an overview of our framework. We use Bonnet [MS19] to train a model on the Bonn sugar beet dataset [CLS*17]. 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 design novel gradient based approaches to select samples, one that is based on the norm of the network gradients, and the other on the norm of projected out gradients. In the experiments section we illustrate the performance of these methods and compare them to other approaches, including random selection of samples, selecting samples driven by the training loss, or driven by the uncertainty of the network.

To compute the measures explained in the following sections, we use what we refer to in the text as pseudo ground truth. These are foreground masks generated by clustering with very weak supervision as detailed in the subsequent section. An example is shown in Figure 3.

### 3.1. Setup

We evaluate our different approaches by first training a network on the Bonn dataset then refining it on the Stuttgart and Zurich datasets. To refine the network we pick unlabeled samples in batches of 10 using one of the methods described in this section. Once the sample 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.

For the methods presented in sections 3.2, 3.3, and 3.4, we first obtain foreground masks with 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 segmentation is sufficient for the purpose of selecting images for annotation. This makes our proposed gradient-based approach feasible in practice.

### 3.2. 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 [LGG*17] based on a pseudo ground truth, consisting of a foreground-background segmentation that we achieve using k-means clustering on the RGB channels.

Initially, we use 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 is merely a few seconds for one dataset. In accordance with previously used terminology [ZGL*18], we refer to this step as being very weakly supervised, since it only involves inspecting a single image.

Pseudo ground truth is generated for all unlabeled images by assigning pixels to the selected clusters, and the loss is computed with respect to it. The images are sorted based on this loss in a descending order. Rather than selecting those with the highest loss, we found that the network learns better when presented with diverse samples. The samples are therefore selected on a log-scale space.

We generate a string of numbers, starting from zero, that are spaced evenly on a log scale. These numbers are then used as the indices of the samples to be selected. 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.

Note that the pseudo ground truth is only used to compute the loss but the network weights remain unchanged and are only later updated with the manual annotations of the selected samples.

### 3.3. 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).

$$n_g(\mathbf{x}) = \left\| \nabla_{w_f} \mathcal{L}(\mathbf{x}) \right\|, \tag{1}$$

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.

### 3.4. 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

Table 1: 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% |

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:

$$n_p(\mathbf{x}) = \left\| \mathbf{g_x} - \sum_{i=1}^{S} \frac{\langle \mathbf{g_i}, \mathbf{g_x} \rangle}{\langle \mathbf{g_i}, \mathbf{g_i} \rangle} \mathbf{g_i} \right\|, \tag{2}$$

where $\mathbf{x}$ is the image, $\mathbf{g_i}$ is the gradient of the $i$th 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.

## 4. 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.

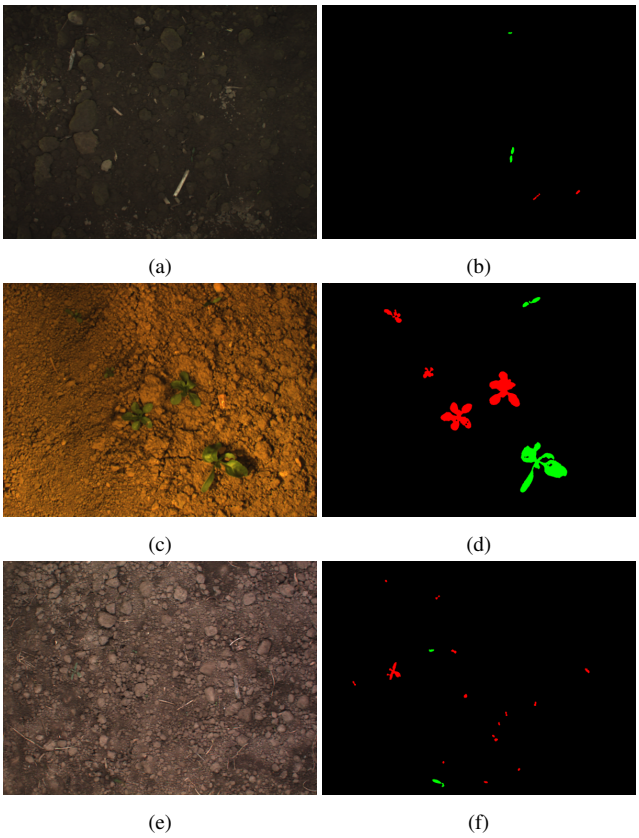
### 4.1. Datasets

The datasets we used were acquired with a Bosch Deepfield Robotics BoniRob UGV 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 2 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: soil, weed and crop. The Bonn dataset is partly publicly available [CLS*17]. Table 1 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 [MLS18] 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.

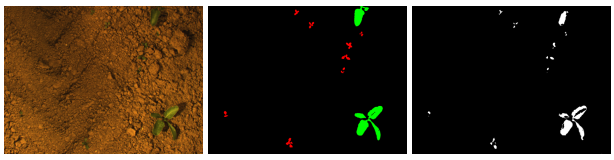
### 4.2. Segmentation Framework

Bonnet is an open-source deep network framework developed by [MS19] that was designed with efficiency in mind. The network is



(a) (b) (c) (d) (e) (f)

Figure 2: Sample images from the Bonn, Stuttgart, and Zurich datasets in the first, second, and third row, respectively. The first column shows the RGB images and the second column shows their annotations (green denotes crop while red denotes weed).



(a) input (b) ground truth (c) pseudo ground truth

Figure 3: Foreground segmentation of vegetation provided by clustering. Note that only a rough segmentation is enough for our approach.

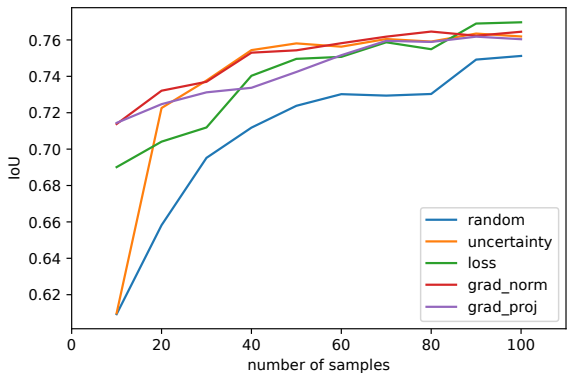



Figure 4: Pixel-wise mean IoU on the Stuttgart dataset. Running the model without any new annotations yields an IoU of 0.34.

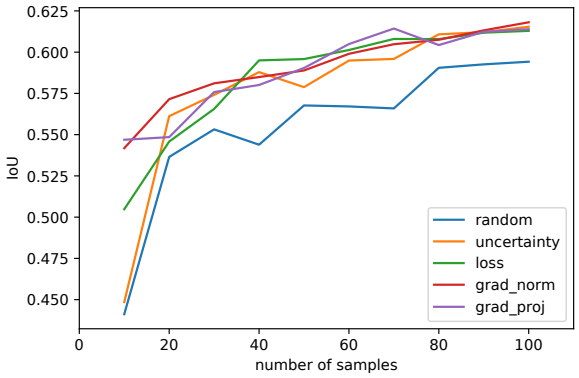



Figure 5: Pixel-wise mean IoU on the Zurich dataset. Running the model without any new annotations yields an IoU of 0.36.

based on SegNet [BKC17] and ENet [PCKC16]. 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 [LGG*17].

To speed up prediction, the authors replace the [5x5] conventional convolutional layer with a mix of [1x1] convolutions and separable [1x5] and [5x1] convolutions. Additionally, instead of using the relatively more expensive transposed convolutions in the decoder, unpooling is done using the respective pooling indices in the encoder part.

As input to our network we only use the RGB channels.

### 4.3. Uncertainty

In addition to the loss and gradient based methods, we also show the performance of an uncertainty driven approach as a baseline. To infer the pixel-wise semantic segmentation of a new image, the network computes softmax probabilities in its last layer. The probabilities can serve as a guide as to which samples the network is most uncertain of. For every image passed through the network, we compute the following measure of the prediction confidence:

$$u(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} \max_c p(c|x_i), \tag{3}$$

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

We then sort the images based on the computed uncertainty measure in a descending order and pick the images accordingly to refine the network on a new dataset. The images are selected on a log-space scale as mentioned in Section 3.

## 5. Results

Figures 4 and 5 show the pixel-wise mean intersection over union (mIoU) on the Stuttgart and Zurich datasets when selecting samples for annotation with different methods. It can be seen from the plots that methods that take into account the impact of the samples on the weights lead to better generalization to the rest of the unseen data, even when presented with a small number of annotated images. In particular, ranking the samples based on the norm of the gradients results in higher mIoU on both datasets.

We also ran an experiment where we trained the model with the pseudo ground truth first and picked samples randomly afterwards. We found that it performs worse than when picking random samples directly. 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.

To further quantify the performance of our approach, we use the object-wise metric defined by [MLS18], 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.

Tables 2 and 3 show how our approach performs on the Stuttgart and Zurich datasets. Each row shows the mean accuracy when selecting $n$ samples with different methods. The baseline is random sampling shown in the first column.

A few observations can be made: the effect of the sampling method is stronger when only a few images are selected. Again, it can be seen that methods measuring the influence of the samples on the weights perform better. For instance, training a model with 20 samples picked with the gradient norm method produces accuracies that can only be achieved when picking 60 samples with the random method, lowering the annotation effort considerably.

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 gradient norm method shows a consistent improvement over other methods for different number of samples and across the two datasets, confirming that samples that might have a larger influence

Table 2: Object-wise Performance on the Stuttgart dataset. 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.

| Samples No. | Random | Uncertainty | Loss | Gradient Norm | Gradient Proj. |
|---|---|---|---|---|---|
| 10 | 0.6920 | 0.6437 | 0.7882 | 0.8040 | 0.8196 |
| 20 | 0.7402 | 0.8408 | 0.7769 | 0.8350 | 0.8404 |
| 30 | 0.8138 | 0.8359 | 0.7950 | 0.8461 | 0.8470 |
| 40 | 0.8254 | 0.8529 | 0.8555 | 0.8682 | 0.8252 |
| 50 | 0.8225 | 0.8529 | 0.8523 | 0.8599 | 0.8278 |
| 60 | 0.8308 | 0.8497 | 0.8596 | 0.8569 | 0.8384 |
| 70 | 0.8335 | 0.8542 | 0.8666 | 0.8622 | 0.8366 |
| 80 | 0.8321 | 0.8595 | 0.8455 | 0.8596 | 0.8386 |
| 90 | 0.8424 | 0.8565 | 0.8643 | 0.8639 | 0.8399 |
| 100 | 0.8394 | 0.8502 | 0.8638 | 0.8531 | 0.8529 |

Table 3: Object-wise Performance on the Zurich dataset. 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 0.33.

| Samples No. | Random | Uncertainty | Loss | Gradient Norm | Gradient Proj. |
|---|---|---|---|---|---|
| 10 | 0.7552 | 0.6943 | 0.7697 | 0.8354 | 0.8025 |
| 20 | 0.7971 | 0.8281 | 0.8189 | 0.8768 | 0.8170 |
| 30 | 0.8591 | 0.8674 | 0.8321 | 0.8553 | 0.8299 |
| 40 | 0.8575 | 0.8690 | 0.8610 | 0.8711 | 0.8479 |
| 50 | 0.8593 | 0.8547 | 0.8636 | 0.8852 | 0.8784 |
| 60 | 0.8666 | 0.8737 | 0.8805 | 0.8827 | 0.8895 |
| 70 | 0.8601 | 0.8664 | 0.8880 | 0.8827 | 0.8878 |
| 80 | 0.8241 | 0.8869 | 0.8867 | 0.8897 | 0.8784 |
| 90 | 0.8476 | 0.8667 | 0.8812 | 0.8928 | 0.8871 |
| 100 | 0.7911 | 0.8700 | 0.8873 | 0.8873 | 0.8805 |

on the weights are more valuable for annotation, as the network can benefit more from them.

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. To further analyze them, 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. We found that both gradient and loss based approaches favor samples that have gradients clustered at the bottom of the plot. In our future work we plan to investigate this further.

A more detailed breakdown of the methods performance is shown in Table 4. 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. The uncertainty-based method, on the other hand, shows a large imbalance of performance on the two classes.

Table 4: 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 are in bold and the lowest in italics. The uncertainty-based method shows a large imbalance in performance on the two classes.

| | Precision | | Recall | |
|---|---|---|---|---|
| | Weed | Crop | Weed | Crop |
| Random | *0.4095* | 0.7278 | 0.4851 | *0.6946* |
| Uncertainty | 0.5580 | *0.6646* | *0.2711* | **0.8880** |
| Loss | 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* |
| Uncertainty | **0.9476** | *0.4586* | *0.4919* | **0.8854** |
| Loss | 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 |

## 6. Conclusion

In this paper, we proposed an active learning approach that supports semantic segmentation in new environments by effectively selecting samples for user annotation with the goal of minimizing the annotation effort. We applied our approach in the domain of crop/weed classification for agricultural robots, reducing the an-

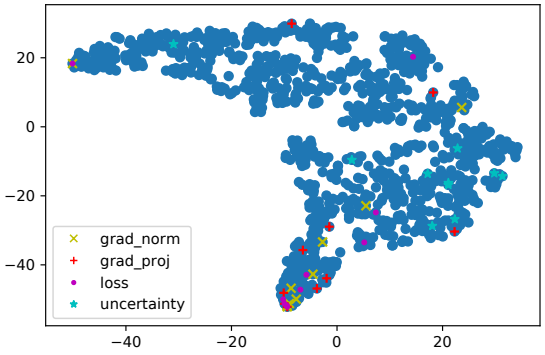



Figure 6: t-SNE of the images gradients on the Stuttgart dataset. Each point represents the 2-D embedding of the gradient vector. The first 10 samples selected by each method are shown in different colors.

notation efforts when moving to different fields or environmental conditions.

In our approach, 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 approach 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 smaller number of training samples, compared to random sampling as well as uncertainty-based approaches for selecting samples for annotation. As a result of that, the effort in human annotation is considerably reduced without compromising performance.

## References

[ALKF18] ACUNA D., LING H., KAR A., FIDLER S.: Efficient interactive annotation of segmentation datasets with polygon-rnn++. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 859–868. 2

[BKC17] BADRINARAYANAN V., KENDALL A., CIPOLLA R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. on Pattern Analalysis and Machine Intelligence (TPAMI) 39*, 12 (2017), 2481–2495. 5

[CLS*17] CHEBROLU N., LOTTES P., SCHAEFER A., WINTERHALTER W., BURGARD W., STACHNISS C.: Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields. *The Intl. Journal of Robotics Research 36*, 10 (2017), 1045–1052. 3, 4

[DCJ*18] DU Y., CZARNECKI W. M., JAYAKUMAR S. M., PASCANU R., LAKSHMINARAYANAN B.: Adapting auxiliary losses using gradient similarity. *arXiv preprint arXiv:1812.02224* (2018). 2

[DJG16] DUTT JAIN S., GRAUMAN K.: Active image segmentation propagation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* (2016), pp. 2864–2873. 2

[DPBG18] DUCKETT T., PEARSON S., BLACKMORE S., GRIEVE B.: Agricultural robotics: The future of robotic agriculture. *arXiv preprint abs/1806.06762* (2018). URL: http://arxiv.org/abs/1806.06762, arXiv:1806.06762. 2

[GCDL11] GUYON I., CAWLEY G., DROR G., LEMAIRE V.: Results of the active learning challenge. In *Proc. of the AISTATS Active Learning and Experimental Design Workshop* (2011), pp. 19–45. 2

[GIG17] GAL Y., ISLAM R., GHAHRAMANI Z.: Deep bayesian active learning with image data. In *Proc. of thhe Intl. Conf. on Machine Learning* (2017), pp. 1183–1192. 2

[HPB08] HOLUB A., PERONA P., BURL M. C.: Entropy-based active learning for object recognition. In *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops* (2008), pp. 1–8. 2

[KHH17] KWAK S., HONG S., HAN B.: Weakly supervised semantic segmentation using superpixel pooling network. In *Thirty-First AAAI Conference on Artificial Intelligence* (2017). 2

[KRFD16] KÄDING C., RODNER E., FREYTAG A., DENZLER J.: Active and continuous exploration with deep neural networks and expected model output changes. *arXiv preprint arXiv:1612.06129* (2016). 2

[LBC*18] LOTTES P., BEHLEY J., CHEBROLU N., MILIOTO A., STACHNISS C.: Joint Stem Detection and Crop-Weed Classification for Plant-specific Treatment in Precision Farming. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)* (2018). 2

[LGG*17] LIN T., GOYAL P., GIRSHICK R. B., HE K., DOLLÁR P.: Focal loss for dense object detection. In *Proc. of the IEEE Intl. Conf. on Computer Vision (ICCV)* (2017), pp. 2999–3007. 3, 5

[LJK*16] LIEBISCH F., JOHANNES P., KHANNA R., LOTTES P., STACHNISS C., FALCK T., SANDER S., SIEGWART R., WALTER A., GALCERAN E.: 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). 2

[MBF*18] MCCOOL C., BEATTIE J., FIRN J., LEHNERT C., KULK J., RUSSELL R., PEREZ T., BAWDEN O.: Efficacy of mechanical weeding tools: A study into alternative weed management strategies enabled by robotics. *IEEE Robotics and Automation Letters (RA-L)* (2018). 2

[MLS17] MILIOTO A., LOTTES P., STACHNISS C.: 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). URL: http://www.ipb.uni-bonn.de/pdfs/milioto17uavg.pdf. 2

[MLS18] MILIOTO A., LOTTES P., STACHNISS C.: 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)* (2018), pp. 2229–2235. 2, 4, 5

[MPU17] MCCOOL C., PEREZ T., UPCROFT B.: Mixtures of Lightweight Deep Convolutional Neural Networks: Applied to Agricultural Robotics. *IEEE Robotics and Automation Letters (RA-L)* (2017). 2

[MS19] MILIOTO A., STACHNISS C.: 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). 3, 4

[PCKC16] PASZKE A., CHAURASIA A., KIM S., CULURCIELLO E.: Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147* (2016). 5

[Set09] SETTLES B.: *Active learning literature survey*. Tech. rep., Univ. of Wisconsin-Madison, Dep. of Computer Sciences, 2009. 2

[SPK*18] SA I., POPOVIC M., KHANNA R., CHEN Z., LOTTES P., LIEBISCH F., NIETO J., STACHNISS C., WALTER A., SIEGWART R.: WeedMap: A Large-Scale Semantic Weed Mapping Framework Using Aerial Multispectral Imaging and Deep Neural Network for Precision Farming. *Remote Sensing 10* (2018). URL: http://www.mdpi.com/2072-4292/10/9/1423/pdf, doi:10.3390/rs10091423. 2

[SS17] SENER O., SAVARESE S.: A geometric approach to active learning for convolutional neural networks. *arXiv preprint arXiv 1708* (2017), 1. 2

[TDP*18] TANG M., DJELOUAH A., PERAZZI F., BOYKOV Y., SCHROERS C.: Normalized cut loss for weakly-supervised cnn segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 1818–1827. 2

[WXS*18] WEI Y., XIAO H., SHI H., JIE Z., FENG J., HUANG T. S.: Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 7268–7277. 2

[WZL*17] WANG K., ZHANG D., LI Y., ZHANG R., LIN L.: Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology 27*, 12 (2017), 2591–2600. 2

[YCBL14] YOSINSKI J., CLUNE J., BENGIO Y., LIPSON H.: How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems 27* (2014), Ghahramani Z., Welling M., Cortes C., Lawrence N. D., Weinberger K. Q., (Eds.), pp. 3320–3328. 1

[YZC*17] YANG L., ZHANG Y., CHEN J., ZHANG S., CHEN D. Z.: 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* (2017), pp. 399–407. 2

[ZGL*18] ZHANG L., GOPALAKRISHNAN V., LU L., SUMMERS R. M., MOSS J., YAO J.: Self-learning to detect and segment cysts in lung ct images without manual annotation. In *IEEE Intl. Symposium on Biomedical Imaging (ISBI 2018)* (2018), pp. 1100–1103. 2, 3

[ZSZ*17] ZHOU Z., SHIN J., ZHANG L., GURUDU S., GOTWAY M., LIANG J.: 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)* (2017), pp. 7340–7351. 2