Calibrated Weed Mapping

Pasquale De Marinis, Gabriele Detomaso, Gennaro Vessio and Giovanna Castellano

Department of Computer Science, University of Bari Aldo Moro, Bari, Italy

Abstract

In the agricultural domain, semantic segmentation models are increasingly used to detect the presence of weeds, thereby enhancing the efficiency of field weeding operations. However, much of the existing research focuses primarily on maximizing predictive accuracy, often overlooking the calibration of model outputs—the alignment between predicted confidence scores and the actual likelihood of correctness. Poor calibration can lead to suboptimal decision-making, resulting in the inefficient use of resources, including excessive herbicide application and increased environmental impact. To address this limitation, we investigate the application of two post-hoc calibration techniques across multiple configurations of two state-of-the-art lightweight segmentation models. The results demonstrate that model calibration enhances the reliability of predictions and provides a viable and effective strategy for improving the practical utility of semantic segmentation in precision agriculture. Code is available at https://github.com/pasqualedem/CalibratedWeedMapping.

Keywords

Drone Vision, Agricultural UAV Imaging, Weed Mapping, Sustainable Agriculture, Neural Network Calibration

1. Introduction

Drone vision represents a rapidly advancing application of computer vision, with diverse use cases ranging from crowd analysis to infrastructure monitoring and environmental surveillance [1, 2]. Among these, precision agriculture stands out as a particularly impactful domain [3]. Unmanned aerial vehicles (UAVs) are well-suited for agricultural tasks due to their ability to cover extensive fields while capturing high-resolution imagery efficiently. This enables detailed monitoring of crop health, soil conditions, and irrigation patterns, all while maintaining relatively low operational costs.

A major challenge in agriculture is the presence of weeds, which compete with crops for nutrients, resulting in the unnecessary depletion of often limited resources. By leveraging drone imagery, it becomes possible to identify infested zones in real time and inform agricultural workers, allowing for immediate intervention. Even better, this process could be automated: upon detection, the drone could trigger herbicide sprayers directly, reducing manual labor and improving efficiency.

Recent studies have demonstrated the potential of deep learning models for semantic segmentation and weed detection in aerial or drone-acquired imagery [4, 5, 6]. Nonetheless, reliable automated weed identification remains a difficult task. Adoption in agriculture is limited, primarily due to the high cost of manual annotation required to generate the large training datasets on which these models depend. This is especially demanding in agricultural contexts, where labeling individual plants in field images is a labor-intensive task. Additionally, the computational demands of these models are incompatible with the limited processing power and energy constraints of drone platforms. These factors, combined with the necessity for real-time operation, highlight the need for more efficient and lightweight approaches to weed mapping.

Although achieving entirely accurate weed mapping predictions may remain elusive, ensuring that model confidence scores reliably reflect predictive accuracy is feasible. This issue has been explored in recent studies on classifier calibration [7]. A calibrated classifier outputs probability estimates that correspond to the actual likelihood of correct classification, based on ground truth data. For example, if

2nd Workshop on Green-Aware Artificial Intelligence, 28th European Conference on Artificial Intelligence (ECAI 2025), October 25–30, 2025, Bologna, Italy

pasquale.demarinis@uniba.it (P. D. Marinis); gennaro.vessio@uniba.it (G. Vessio); giovanna.castellano@uniba.it (G. Castellano)

0000-0001-8935-9156 (P. D. Marinis); 0000-0002-0883-2691 (G. Vessio); 0000-0002-6489-8628 (G. Castellano)

a model assigns a confidence score of 0.4 to 100 instances predicted as "dog", then approximately 40 of those instances should actually be labeled as "dog" [8].

A well-calibrated model could significantly improve decision-making in weed mapping, whose primary objectives are to enhance agricultural productivity and reduce herbicide usage. By providing reliable uncertainty estimates, such a model would enable farmers to make better-informed choices about when and where to apply weed treatments. For instance, if the model exhibits high confidence regarding the presence of weeds in a specific area—even when the background class has higher overall confidence—a farmer might still opt to apply herbicide in that location, relying on the model's localized prediction. Conversely, if the objective is to minimize herbicide usage, treatment could be restricted to areas where the model's confidence in weed presence surpasses a predefined threshold, thereby reducing unnecessary chemical application in regions likely to be weed-free. This targeted approach increases operational efficiency, lowers environmental impact, and promotes more sustainable agricultural practices, illustrating how calibration in AI models contributes to *green-aware* AI.

In this paper, we explore the calibration problem in weed mapping, focusing on two lightweight semantic segmentation models: SegFormer [9] and MobileNetV4 [10]. We evaluate the impact of different calibration techniques on these models, specifically matrix scaling [7] and temperature scaling [7], to determine their effectiveness in improving model reliability. The WeedMap dataset [11] is used for training and evaluation.

2. Related Work

Several studies have already been conducted in the field of weed mapping. WeedNet, which utilizes the SegNet architecture, represents a practical implementation in this field [5]. The WeedMap dataset, composed of multispectral images collected from sugar beet fields in Germany and Switzerland, serves as a recognized benchmark for research on weed mapping [11]. In their work, the authors trained a semantic segmentation model to distinguish weeds within crop fields and explored various channel combinations to identify those most discriminative for the segmentation task. The multispectral issue and its adoption have been further investigated in [6], where we propose different methods to reuse RGB pre-training on multispectral input data. In [12], instead, we tackle the resource constraint problem using knowledge distillation.

Partially supervised techniques, including semi-supervised and unsupervised learning, have been investigated for weed detection. Semi-supervised approaches are detailed in works such as [13, 14, 15]. Crop-row detection techniques, notably those utilizing the Hough Transform, have been applied to differentiate crops from weeds [16, 17, 18, 19, 20, 21, 22]. These studies are a foundational basis for training accurate and robust models, although not necessarily calibrated ones.

Calibration of deep learning models has been extensively studied in various domains, including image classification [23], object detection [24], and semantic segmentation [8, 25]. These works have explored various calibration techniques, including temperature scaling, matrix scaling, and label smoothing, to enhance the reliability of model predictions. However, the application of these techniques to weed mapping remains largely unexplored.

3. Materials

The WeedMap dataset [11] was used for this study. The original data includes eight sugar beet fields: five in Rheinbach, Germany, and three in Eschikon, Switzerland. Each field has multiple images available—one for each channel, along with the ground truth—resulting in 12 image channels plus ground truth for the German fields and 8 channels plus ground truth for the Swiss fields. In particular, this study utilized the RGB subset of the original dataset, which has been previously used in other works. The goal was not to improve classification accuracy but rather to assess and enhance model calibration.

Specifically, five images (RGB channel only), numbered from 000 to 004 and corresponding to the German fields, were selected along with their respective ground truth annotations. These images were then rotated to align the crop rows horizontally. Given that the selected semantic segmentation models accept only inputs of size 512×512, fixed-size patches were extracted from the original images and ground truths to conform to this input requirement.

The final dataset comprises 352 image patches paired with their corresponding ground truth masks. The dataset was divided as follows: patches from images 000, 002, and 004 were used for training (236 patches); patches from image 001 were used for validation (52 patches); and patches from image 003 were reserved for testing (65 patches).

4. Methods

The proposed framework consists of three main components: (1) a lightweight semantic segmentation model, (2) a calibration technique, and (3) an evaluation phase. The semantic segmentation model is trained on the WeedMap dataset to segment weeds, crops, and background. The calibration technique is applied to the model's predictions to improve reliability using the validation set. Finally, the evaluation phase assesses the model's performance using metrics such as F1 score, Expected Calibration Error (ECE), Static Calibration Error (SCE), and reliability diagrams.

4.1. Model Training

For this study, we selected two lightweight models: SegFormer-B0 [9], which has also been adopted in prior work on weed mapping [21], and the newer MobileNetV4 [10].

SegFormer's architecture consists of two main components: (1) a hierarchical Transformer encoder that captures both coarse, high-resolution features and fine, low-resolution features; and (2) a lightweight MLP decoder that fuses these multi-scale features to produce the final semantic segmentation mask. By adjusting the dimensions of the feature maps at various stages of the encoder, different SegFormer configurations can be obtained. Specifically, the MiT-B0 model [9] is well-suited for scenarios like ours, where low computational overhead and near real-time inference are critical constraints.

MobileNetV4 is a lightweight model combining depthwise separable convolutions and inverted residual blocks to achieve high accuracy and low computational cost. It is designed to be efficient in terms of both memory and processing power, making it suitable for deployment on resource-constrained devices, such as drones. Since the model does not include a decoder for semantic segmentation, we employed the pre-trained classification checkpoint released by the authors and integrated it with the lightweight SegFormer decoder.

The model encoders were initially imported pre-trained on the ImageNet-1k dataset [26]. However, an additional fine-tuning phase was performed on the WeedMap training set to specialize them for segmenting images containing the three target classes: background, crop, and weed.

For each model, two configurations were designed: the first one was trained using the standard cross-entropy loss. The second one employed the focal loss, chosen because it shares similarities with the cross-entropy loss but places greater emphasis on training examples classified with high uncertainty [27], thus addressing dataset imbalance. Moreover, it is theoretically expected that applying calibration techniques to the focal loss model should yield better results than cross-entropy [28]. The focal loss function is defined as:

$$\mathscr{L}_{FL}(p_t) = -\alpha_t(1 - p_t)^{\gamma} \log(p_t)$$

where p_t is the model's predicted probability for the true class, α_t is a weighting factor for the class, and γ is a focusing parameter that adjusts the rate at which easy examples are down-weighted. The hyperparameter γ was not fixed but learned during training after initialization.

4.2. Calibration

We apply two post-hoc techniques (after training) to address the central problem of calibration: matrix scaling and temperature scaling [7]. These methods were selected for their widespread use and well-established properties, providing a reliable baseline for the first exploration of calibration in weed mapping.

Matrix scaling uses the predicted class scores (logits) as input to learning a logistic regressor that outputs transformed logits:

$$(Wz + b) \in \mathbb{R}^k$$

where $z \in \mathbb{R}^k$ is the logit vector for a pixel from the last layer, while $W \in \mathbb{R}^{k \times k}$ and $b \in \mathbb{R}^k$ are the calibration parameters to be learned. Using these transformed logits, calibrated probabilities and predicted classes are obtained as follows:

$$\hat{q} = \max_k \text{softmax}(Wz + b)^{(k)}$$
$$\hat{y} = \arg\max_k (Wz + b)^{(k)}$$

where $\hat{q}$ is the probability of the predicted class and $\hat{y}$ is the predicted class. In this way, the logits are transformed into calibrated probabilities that can be interpreted as confidence scores.

Temperature scaling rescales the logits using a temperature parameter T so that the softmax applied to the scaled logits yields calibrated probabilities:

$$\text{softmax}\left(\frac{\boldsymbol{z}}{T}\right)$$

Values of $T < 1$ sharpen the probability distribution (increasing confidence), while $T > 1$ smooths the distribution, increasing entropy. Calibrated probabilities and predicted classes are computed as:

$$\hat{q} = \max_k \text{softmax}\left(\frac{\boldsymbol{z}}{T}\right)^{(k)}$$
$$\hat{y} = \arg\max_k \left(\frac{\boldsymbol{z}}{T}\right)^{(k)}$$

where $\hat{q}$ is the probability of the predicted class and $\hat{y}$ is the predicted class.

The models are then fine-tuned on the validation set using the cross-entropy loss for both techniques. However, since both methods are post-hoc, only the calibration parameters were updated: W and b for matrix scaling and T for temperature scaling.

4.3. Evaluation

For the evaluation on the test set, three metrics were calculated: (1) the F1 score to assess whether the calibration techniques affected the model's accuracy, (2) the Expected Calibration Error (ECE) to evaluate the various calibrations, and (3) the Static Calibration Error (SCE) to provide additional confirmation of the calibration results. In addition, we calculated the reliability diagram for each configuration to represent miscalibration visually.

Expected Calibration Error [29] is a scalar value that quantifies the model's miscalibration. It is the weighted average difference between accuracy and confidence across M bins for N instances:

$$\text{ECE} = \sum_{b=1}^{M} \frac{|Bin_b|}{N} \left|\text{acc}(Bin_b) - \text{conf}(Bin_b)\right|$$

Perfect calibration would require ECE $= 0$, but achieving a perfectly calibrated model is impossible [30].

Although this metric is widely used, it may not fully capture miscalibration in multi-class classification tasks. Therefore, to obtain additional confirmation, the Static Calibration Error was also computed, which is more robust as it accounts for class imbalance by computing the error per class and averaging:

$$\mathrm{SCE} = \frac{1}{K}\sum_{k=1}^{K}\sum_{b=1}^{M}\frac{|Bin_{bk}|}{N}\,|\mathrm{acc}(Bin_{bk}) - \mathrm{conf}(Bin_{bk})|$$

where K is the number of classes and Bin_{bk} is the b-th bin containing pixels whose ground truth is k.

5. Experiments

Both SegFormer and MobileNetV4 were trained for 30 epochs using the AdamW optimizer with a learning rate of 0.00006. For matrix scaling, three fine-tuning durations (15, 50, and 100 epochs) were tested on the base models to evaluate the trade-off between computation time and performance. AdamW was reused with a learning rate of 0.006. For temperature scaling, two configurations (15 and 30 epochs) were evaluated, as this method typically converges faster than matrix scaling. The same optimizer and learning rate (AdamW, 0.006) were used. We report only the results corresponding to the highest number of epochs in each configuration since they consistently outperformed the shorter training runs.

When using focal loss, the models were trained with the same settings as the baselines, except for the loss function. In this case, the focal loss was applied, introducing the hyperparameter γ, which was not fixed but learned during training from an initial value of 2.0.

All model training sessions were conducted on an RTX 4090 GPU with 24 GB of memory.

5.1. Comparison of Different Techniques

As shown in Table 1, the baseline models trained with the cross-entropy loss achieved a satisfactory average F1 score. However, their performance deteriorates markedly in terms of accuracy, particularly for the weed class. This decline is a well-documented challenge in weed mapping tasks, as the weed class is typically underrepresented in training datasets. In contrast, models trained with the focal loss consistently outperformed those using cross-entropy across all configurations, with notable improvements in the classification accuracy of the weed class. This suggests that the focal loss is more effective in addressing class imbalance by emphasizing harder-to-classify examples. Additionally, the reliability diagrams in Figs. 1a, 1d, 2a, and 2d indicate that both cross-entropy and focal loss-based models exhibit slight miscalibration, with cross-entropy models being more affected. Specifically, both sets of models tend to be under-confident—often assigning low probability scores even when the prediction is correct.

The F1 scores reported in Table 1 for models using matrix scaling did not reach the levels achieved by the corresponding base models. This discrepancy could represent a limitation in the context of

Table 1
F1 scores (%) for each architecture, method, and loss function. CE stands for cross-entropy, and FL for focal loss. Temperature scaling is not shown, as it does not affect predictions; therefore, F1 scores remain identical to those of the base model.

Architecture	Method	Loss	Avg. F1	Background	Crop	Weed
SegFormer	Base model	CE	80.86	98.60	79.46	64.51
		FL	**81.91**	**98.63**	**80.29**	**66.81**
	Matrix scaling	CE	75.83	98.51	73.21	55.78
		FL	71.93	98.40	70.89	46.50
MobileNetV4	Base model	CE	82.65	98.67	81.90	67.38
		FL	**82.94**	**98.69**	**82.41**	**67.72**
	Matrix scaling	CE	81.71	98.80	82.88	63.45
		FL	79.68	98.67	75.28	65.09

Table 2
Expected Calibration Error (ECE) and Static Calibration Error (SCE) for each architecture, method, and loss function.

Architecture	Method	Loss	ECE	SCE
SegFormer	Base model	CE	0.04898	0.16631
		FL	0.11312	0.21129
	Matrix scaling	CE	0.01248	0.16338
		FL	0.01652	0.16582
	Temperature scaling	CE	**0.00300**	**0.09574**
		FL	0.00602	0.10021
MobileNetV4	Base model	CE	0.05722	0.15726
		FL	0.12284	0.21919
	Matrix scaling	CE	0.03002	0.15921
		FL	0.03383	0.18044
	Temperature scaling	CE	**0.00802**	**0.10226**
		FL	0.01343	0.11287

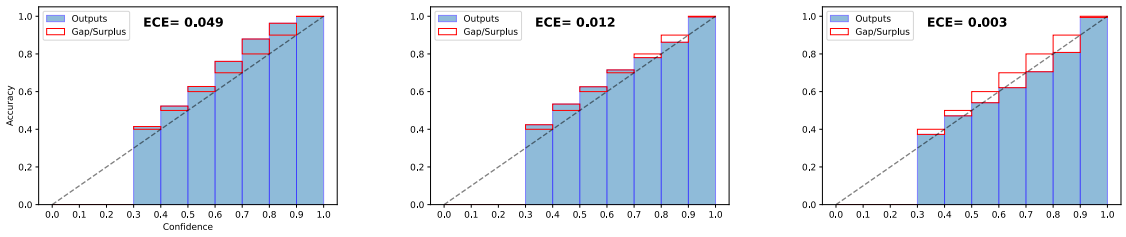
Table 3
Optimal temperature values T learned during temperature scaling for each architecture and loss function.

Configuration	SegFormer-CE	SegFormer-FL	MobileNetV4-CE	MobileNetV4-FL
Temperature T	0.458	0.462	0.652	0.493

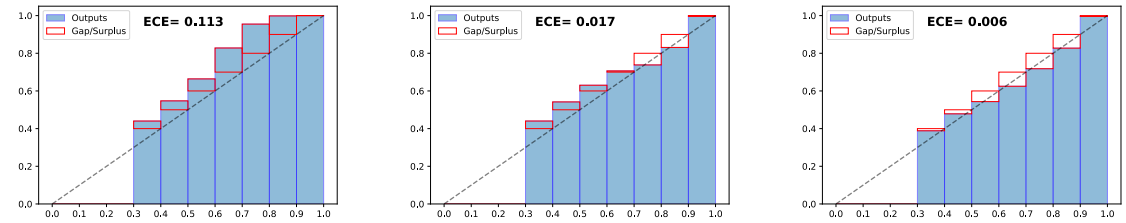
precision agriculture, where the primary goal is the accurate detection of weeds. The focal loss is a promising approach to counteract this limitation. Notably, the MobileNetV4 model trained with the focal loss achieves a classification accuracy for the challenging weed class that closely approaches that of the corresponding base model despite exhibiting a slightly lower overall F1 score. This outcome is in line with the intended behavior of focal loss.

Regarding the primary objective of this study—model calibration—all models with matrix scaling yielded good results. As shown in Table 2, all SCE scores have either dropped or remained unchanged. For temperature scaling, it is observed that a brief fine-tuning of the temperature parameter is sufficient to match the performance of the baseline models. The reliability diagrams in Figs. 1c, 1f, 2c and 2f indicate excellent calibration, although the calibrated models exhibit a slight degree of overconfidence. Calibration leaves performance unchanged, since temperature scaling only modifies the confidence of predictions without altering the predicted class. In this case, as shown in Table 3, the optimized temperature is less than 1 (i.e., a "cold" temperature), which sharpens the predicted class probability distributions by reducing entropy. Given that the baseline models tend to be under-confident, this sharpening increases the predicted confidence without altering the predicted class labels.

Calibration with temperature scaling overcomes all the configurations. Considering only the reliability diagrams of the models fine-tuned with temperature scaling in Figs. 1 and 2, the apparent small gap—going from one loss to the other—visually suggests a substantial improvement. However, both ECE and SCE scores indicate the opposite (cross-entropy has the best SCE score for both SegFormer and MobileNetV4). The explanation for the discrepancy is as follows: bins exhibiting a large gap contain only a small number of pixels classified with those confidence levels; in fact, temperature scaling has pushed many confidences upward. Therefore, during the ECE calculation, the gap's importance is diminished because it is weighted by the factor $\frac{N_b}{N}$ where N_b is the number of pixels classified with confidence values and N is the total number of pixels. Hence, the ECE score of the gap makes a small contribution to the total ECE score.

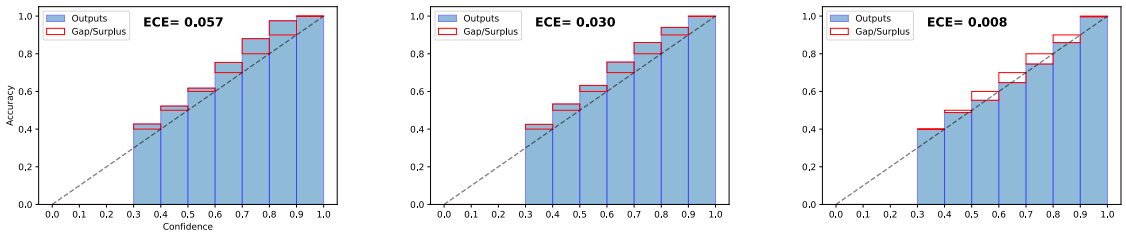


(a) Base model (CE) (b) Matrix scaling (CE) (c) Temperature scaling (CE)

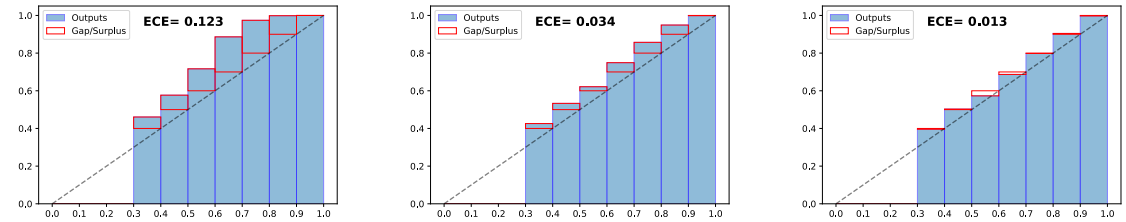


(d) Base model (FL) (e) Matrix scaling (FL) (f) Temperature scaling (FL)

Figure 1: Reliability diagrams for each method with SegFormer.



(a) Base model (CE) (b) Matrix scaling (CE) (c) Temperature scaling (CE)



(d) Base model (FL) (e) Matrix scaling (FL) (f) Temperature scaling (FL)

Figure 2: Reliability diagrams for each method with MobileNetV4.

5.2. Qualitative Evaluation

Figure 3 presents a qualitative comparison of the weed-class probability maps produced by the uncalibrated SegFormer model (base model) and the same model after calibration using temperature scaling. Specifically, areas with false negatives for the weed class are marked by white circles. This example illustrates a practical application of the calibration techniques discussed in this work.

The uncalibrated model demonstrates underconfidence and low recall for the weed class—an issue often observed in weed mapping due to class imbalance and visual ambiguity. While the segmentation outputs of both models appear identical, their corresponding probability maps reveal substantial differences. Calibration at a lower temperature increases the divergence of the probability distribution from uniformity, yielding higher model confidence. Specifically, in regions where weed instances are missed (false negatives), the calibrated model assigns a higher probability to the background class. At

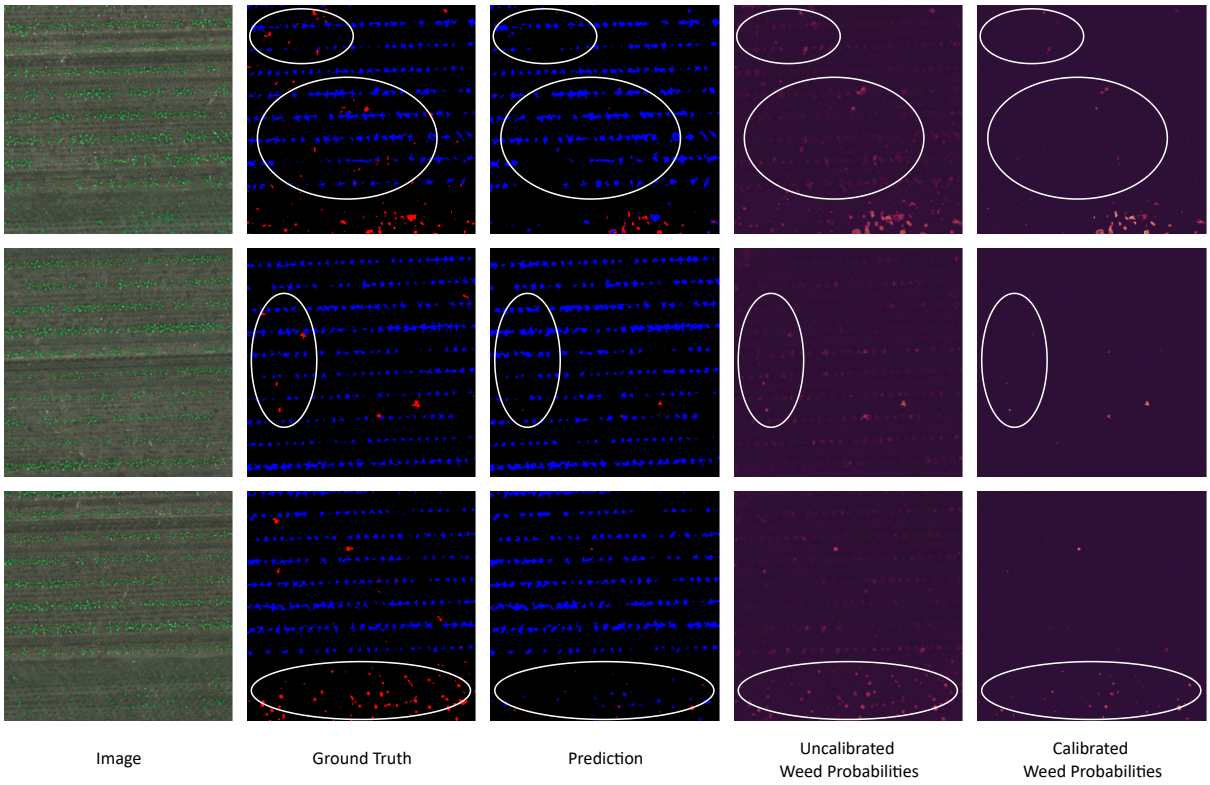


Figure 3: Examples of calibration results visualized on the probabilities of the weed class obtained with the uncalibrated SegFormer model (base model) and the calibrated model with temperature scaling. Circle-marked areas indicate where the model is unable to recognize weeds, but the calibrated model can assign a higher probability of weed presence.

the same time, it increases the separation between the probabilities assigned to crop and weed classes, thereby enhancing the interpretability and reliability of the probability map.

In practical deployments, this distinction is critical. Even in areas where the model fails to predict the presence of weeds explicitly (circled areas in Fig. 3), a threshold-based decision system operating on calibrated probabilities can still detect weed presence. This capability is essential in precision agriculture, where accurate confidence estimates enable more efficient resource allocation and help minimize herbicide usage.

6. Conclusion

The objective of this research was to experiment with two calibration techniques to enhance model reliability, enabling farmers to make informed decisions and prioritize issues based on the confidence reported by the model.

Focal loss improves classification accuracy, particularly for classes that are underrepresented or inherently ambiguous. However, it consistently produces lower probabilistic calibration compared to cross-entropy, reflecting the tendency of focal loss to overemphasize hard-to-classify samples. Applying temperature scaling effectively mitigates this miscalibration, improving the reliability of models trained with focal loss. Despite this improvement, models trained with cross-entropy combined with temperature scaling achieve the best overall calibration, even if their classification accuracy is slightly lower. Taken together, these results suggest that while focal loss can provide modest gains in accuracy for challenging classes, cross-entropy with post-hoc calibration offers the most reliable balance between predictive performance and confidence estimation. From the perspective of automating (even partially) agricultural operations, improving model calibration—particularly for weed detection—would enable agricultural workers to rely entirely on the system for weed removal, prioritizing areas to treat based

on available resources and dedicating their attention only to minor refinements. By enabling targeted herbicide application and optimizing UAV resource use, calibrated models improve operational efficiency while contributing to environmentally sustainable, green-aware AI practices.

Further investigations to continue this research include: (1) refining the fine-tuning of the matrix scaling technique to identify a configuration that calibrates the model without degrading (or ideally improving) the F1 score, (2) exploring additional calibration methods, such as label smoothing, and (3) evaluating the approach on larger or alternative datasets with comprehensive ground truth annotations, which would enable a more thorough assessment of model calibration in diverse agricultural contexts.

Acknowledgments

The research of P. De Marinis is supported by a Ph.D. fellowship funded under the Italian "D.M. n. 352, April 9, 2022" – NRRP, Mission 4, Component 2, Investment 3.3 – co-supported by Exprivia S.p.A. (CUP H91I22000410007).

Declaration on Generative AI

While preparing this work, the authors utilized ChatGPT and Grammarly to enhance language clarity and readability. The authors, who take full responsibility for the final version of the manuscript, carefully reviewed and refined all content generated by these tools.

References

[1] A. Varghese, J. Gubbi, H. Sharma, P. Balamuralidhar, Power infrastructure monitoring and damage detection using drone captured images, in: 2017 International Joint Conference on Neural Networks (IJCNN), 2017, pp. 1681–1687.

[2] G. Castellano, E. Cotardo, C. Mencar, G. Vessio, Density-based clustering with fully-convolutional networks for crowd flow detection from drones, Neurocomput. 526 (2023) 169–179.

[3] P. Daponte, L. De Vito, L. Glielmo, L. Iannelli, D. Liuzza, F. Picariello, G. Silano, A review on the use of drones for precision agriculture, IOP Conference Series: Earth and Environmental Science 275 (2019) 012022. Publisher: IOP Publishing.

[4] A. dos Santos Ferreira, D. M. Freitas, G. G. da Silva, H. Pistori, M. T. Folhes, Weed Detection in Soybean Crops Using ConvNets, Computers and Electronics in Agriculture 143 (2017) 314–324.

[5] I. Sa, Z. Chen, M. Popović, R. Khanna, F. Liebisch, J. Nieto, R. Siegwart, Weednet: Dense Semantic Weed Classification Using Multispectral Images and Mav for Smart Farming, IEEE robotics and automation letters 3 (2017) 588–595.

[6] G. Castellano, P. De Marinis, G. Vessio, Weed mapping in multispectral drone imagery using lightweight vision transformers, Neurocomputing 562 (2023) 126914.

[7] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On Calibration of Modern Neural Networks, in: Proceedings of the 34th International Conference on Machine Learning, PMLR, 2017, pp. 1321–1330. ISSN: 2640-3498.

[8] Z. Lin, S. Trivedi, J. Sun, Taking a Step Back with KCal: Multi-Class Kernel-Based Calibration for Deep Neural Networks (2023).

[9] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, P. Luo, SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers, arXiv preprint arXiv:2105.15203 (2021).

[10] D. Qin, C. Leichner, M. Delakis, M. Fornoni, S. Luo, F. Yang, W. Wang, C. Banbury, C. Ye, B. Akin, V. Aggarwal, T. Zhu, D. Moro, A. Howard, MobileNetV4: Universal Models for the Mobile Ecosystem, in: A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, G. Varol (Eds.), Computer Vision – ECCV 2024, Springer Nature Switzerland, Cham, 2025, pp. 78–96.

[11] I. Sa, M. Popović, R. Khanna, Z. Chen, P. Lottes, F. Liebisch, J. Nieto, C. Stachniss, A. Walter, R. Siegwart, WeedMap: A Large-Scale Semantic Weed Mapping Framework Using Aerial Multispectral

Imaging and Deep Neural Network for Precision Farming, Remote Sensing 10 (2018) 1423. Number: 9 Publisher: Multidisciplinary Digital Publishing Institute.
[12] G. Castellano, P. De Marinis, G. Vessio, Applying Knowledge Distillation to Improve Weed Mapping with Drones, in: 2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS), 2023, pp. 393–400.
[13] C. Nong, X. Fan, J. Wang, Semi-supervised Learning for Weed and Crop Segmentation Using UAV Imagery, Frontiers in Plant Science 13 (2022). Publisher: Frontiers.
[14] S. Khan, M. Tufail, M. T. Khan, Z. A. Khan, J. Iqbal, M. Alam, A novel semi-supervised framework for UAV based crop/weed classification, PLOS ONE 16 (2021) e0251008. Publisher: Public Library of Science.
[15] S. Shorewala, A. Ashfaque, R. Sidharth, U. Verma, Weed Density and Distribution Estimation for Precision Agriculture Using Semi-Supervised Learning, IEEE Access 9 (2021) 27971–27986. Conference Name: IEEE Access.
[16] J. M. Peña, J. Torres-Sánchez, A. I. d. Castro, M. Kelly, F. López-Granados, Weed Mapping in Early-Season Maize Fields Using Object-Based Analysis of Unmanned Aerial Vehicle (UAV) Images, PLOS ONE 8 (2013) e77151. Publisher: Public Library of Science.
[17] A. dos Santos Ferreira, D. M. Freitas, G. G. da Silva, H. Pistori, M. T. Folhes, Unsupervised deep learning and semi-automatic data labeling in weed discrimination, Computers and Electronics in Agriculture 165 (2019) 104963.
[18] M. Pérez-Ortiz, J. M. Peña, P. A. Gutiérrez, J. Torres-Sánchez, C. Hervás-Martínez, F. López-Granados, Selecting patterns and features for between- and within- crop-row weed mapping using UAV-imagery, Expert Systems with Applications 47 (2016) 85–94.
[19] M. D. Bah, A. Hafiane, R. Canals, Deep Learning with Unsupervised Data Labeling for Weed Detection in Line Crops in UAV Images, Remote Sensing 10 (2018) 1690. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.
[20] M. D. Bah, A. Hafiane, R. Canals, B. Emile, Deep features and One-class classification with unsupervised data for weed detection in UAV images, in: 2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA), 2019, pp. 1–5. ISSN: 2154-512X.
[21] P. De Marinis, G. Vessio, G. Castellano, RoWeeder: Unsupervised Weed Mapping Through Crop-Row Detection, in: A. Del Bue, C. Canton, J. Pont-Tuset, T. Tommasi (Eds.), Computer Vision – ECCV 2024 Workshops, Springer Nature Switzerland, Cham, 2025, pp. 132–145.
[22] G. Roggiolani, J. Rückin, M. Popović, J. Behley, C. Stachniss, Unsupervised semantic label generation in agricultural fields, Frontiers in Robotics and AI 12 (2025) 1548143. Publisher: Frontiers.
[23] A. S. Sambyal, U. Niyaz, N. C. Krishnan, D. R. Bathula, Understanding calibration of deep neural networks for medical image classification, Computer Methods and Programs in Biomedicine 242 (2023) 107816.
[24] M. A. Munir, M. H. Khan, M. Sarfraz, M. Ali, Towards Improving Calibration in Object Detection Under Domain Shift, Advances in Neural Information Processing Systems 35 (2022) 38706–38718.
[25] D. Wang, B. Gong, L. Wang, On Calibrating Semantic Segmentation Models: Analyses and An Algorithm, 2023. ArXiv:2212.12053 [cs].
[26] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, in: Advances in Neural Information Processing Systems, volume 25, Curran Associates, Inc., 2012.
[27] T. Lin, P. Goyal, R. B. Girshick, K. He, P. Dollár, Focal Loss for Dense Object Detection, CoRR abs/1708.02002 (2017).
[28] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. Torr, P. Dokania, Calibrating Deep Neural Networks using Focal Loss, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 15288–15299.
[29] M. P. Naeini, G. Cooper, M. Hauskrecht, Obtaining Well Calibrated Probabilities Using Bayesian Binning, Proceedings of the AAAI Conference on Artificial Intelligence 29 (2015). Number: 1.
[30] J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, D. Tran, Measuring Calibration in Deep Learning (2019).