Proceedings of the 18th Conference on Computer Science and Intelligence Systems pp. 393–400
DOI: 10.15439/2023F960
ISSN 2300-5963 ACSIS, Vol. 35

Applying Knowledge Distillation to Improve Weed Mapping With Drones

Giovanna Castellano, Pasquale De Marinis, Gennaro Vessio
0000-0002-6489-8628
0000-0001-8935-9156
0000-0002-0883-2691
Department of Computer Science
University of Bari Aldo Moro
Bari, Italy
Email: {giovanna.castellano, pasquale.demarinis, gennaro.vessio}@uniba.it

***Abstract*—In precision agriculture, non-invasive remote sensing using UAVs can be employed to observe crops in visible and non-visible spectra. This paper investigates the effectiveness of state-of-the-art knowledge distillation techniques for mapping weeds with drones, an essential component of precision agriculture that employs remote sensing to monitor crops and weeds. The study introduces a lightweight Vision Transformer-based model that achieves optimal weed mapping capabilities while maintaining minimal computation time. The research shows that the student model effectively learns from the teacher model using the WeedMap dataset, achieving accurate results suitable for mobile platforms such as drones, with only 0.5 GMacs compared to 42.5 GMacs of the teacher model. The trained models obtained an F1 score of 0.863 and 0.631 on two data subsets, with a performance improvement of 2 and 7 points, respectively, over the undistilled model. The study results suggest that developing efficient computer vision algorithms on drones can significantly improve agricultural management practices, leading to greater profitability and environmental sustainability.**

I. Introduction

PRECISION agriculture has become increasingly important due to the growth of the world's population—which is expected to reach nine billion people by 2050—and the resulting need to increase food production [1]. However, the resources that sustain agriculture are becoming increasingly scarce, degraded, and vulnerable to climate change. This has led to the need for more sustainable and efficient agricultural practices that make optimal use of available resources.

Unmanned aerial vehicles (UAVs), also known as drones, have emerged as valuable tools for precision agriculture due to their versatility and affordability [2]. They can capture high-resolution images and data from agricultural fields, which can be used to monitor crop growth, identify diseases and pests, and optimize irrigation. By providing farmers with accurate and timely information, drones can help reduce costs, increase yields, and minimize the use of inputs such as water, fertilizer, and pesticides. Traditional ground-based methods, such as manual exploration or satellite remote sensing, cannot match the level of detail that drones can provide. Drones can fly over crops and capture images and data in real-time, enabling farmers to make informed decisions quickly. Another benefit of drones is their ability to quickly and efficiently cover large areas. Drones can fly over fields and acquire data in hours, which would otherwise take days or weeks with traditional methods. This can help farmers save time and resources and make more timely decisions. Newer techniques now allow their application in swarm configurations, significantly improving their efficiency and the range of tasks they can perform [3].

One area of particular interest in precision agriculture that could benefit significantly from using drones is weed mapping. Weed mapping is critical in precision agriculture because it allows farmers to apply herbicides accurately and reduce overuse, which can cause environmental and health problems. Convolutional Neural Network (CNN)-based models have been recently proposed to perform semantic segmentation and identify weeds in images captured by drones or other aerial vehicles [4], [5], [6]. However, these models are typically computationally expensive, making them difficult to implement on drones with limited processing power and limited battery. These limitations and the need for real-time responses call for lightweight solutions.

In cases like this, knowledge distillation (KD) can help. KD is a technique that allows a smaller model to learn from a larger, more complex model, often referred to as a *teacher* [7]. The goal is to transfer the knowledge learned from the teacher to a smaller, lighter *student*, which can be used on resource-limited devices such as drones.

In this study, we explore the application of different knowledge distillation techniques to evaluate their effectiveness in the context of drone-based weed mapping. The focus is on using the WeedMap dataset [8] and proposing specific, extra-lightweight architectural designs that aim to achieve superior weed mapping capabilities while maintaining minimal computational time. This technology can improve weed management practices, leading to more sustainable and efficient agriculture.

The rest of this paper is structured as follows. Section II reviews related work. Sections III and IV describe materials and methods. Section V reports and discusses the experimental results. Section VI concludes the paper and outlines future developments in our research.

II. Related Work

Many precision agriculture tasks have been addressed thanks to the recent development of computer vision techniques and remote sensing data collection methodologies. Recent tasks include disease and pest identification, abiotic stress assessment, growth monitoring, crop yield prediction, and weed mapping.

A. Weed Mapping

Weed mapping is a semantic segmentation task in which each pixel of an image is assigned a class. Deep learning algorithms have significantly outperformed more traditional techniques in this task. Dos Santos et al. [4] were among the first to demonstrate the superiority of CNNs, particularly AlexNet, over traditional machine learning approaches such as SVM and Random Forest. Lottes et al. [9] used a CNN with two decoders, one for detecting stem position and the other for plant segmentation. They used the BoniRob dataset and one collected with a UAV for evaluation. They obtained an mAP of 79.2% for stem detection and 75.3% for segmentation. SegNet with ResNet50 as an encoder was used in [6], achieving an F1 score of 64.6%.

Depending on the bands acquired, multispectral images can contain information on the growth and health status of a plant and its species. Consequently, they can improve the accuracy of deep learning models compared to models trained only with RGB. Furthermore, multispectral image sensors can be easily integrated into UAVs. The popular U-Net was used on a dataset available on the Internet to separate weeds from crops and soil, achieving an F1 score of 89% and a mIoU of 98% [10]. WeedNet, a semantic segmentation network based on SegNet, was developed and trained on the WeedMap dataset and achieved an F1 score of 80% [5]. WeedMap contains two sets of images of sugar beet fields collected in Germany (Rheinbach) and Switzerland (Eschikon). Both were collected using UAVs equipped with multispectral cameras. The former used a 5-channel RedEdge-M camera, while the latter used a 4-channel Sequoia camera. The authors also trained SegNet using various combinations of the acquired channels, resulting in an AUC of 84.3% [8]. On the WeedMap dataset, the DeepLabV3 architecture for semantic segmentation was compared with SegNet and U-Net, achieving an F1 score of 81% on the Rheinbach subset [11]. Mozzam et al. [12] used patch-based training with a modified VGG model on the same dataset, also using the Eschikon subset. The patches were chosen by hand, and those that contained both classes were removed. On the Rheinbach subset, the accuracy reached 92%, and on Eschikon 90%. WeedMap has been used here for benchmarking purposes, as it has become the preferred dataset in several works due to its volume and quality.

B. Knowledge Distillation

Knowledge distillation is a popular method for "compressing" neural models [7]. However, previous studies have shown a significant size gap between student and teacher networks, which limits the effectiveness of KD. Mirzadeh et al. [13] showed that the gap between student and teacher could not be arbitrary and proposed a solution for this problem, called *teacher assistant knowledge distillation*. This method requires training one or more teaching assistant networks and is computationally expensive. In addition, errors of the teaching assistant can accumulate and transfer to the student. To mitigate these problems, Jafari et al. [14] introduced *annealing KD*, which achieves state-of-the-art performance in natural language understanding and computer vision tasks. In annealing KD, the teacher's goals are annealed to convey the information provided by the teacher to the student gradually. The predictions are annealed using a temperature parameter that gradually decreases during training. After this first phase, the student is trained with the ground truth. Although it can handle the capability gap problem, annealing KD is still vulnerable to noisy data and teachers' results. In addition, the training requires deciding when to switch from the first to the second phase, which can be challenging. Inspired by continuation optimization, Jafari et al. [15] tried to solve the above problems by introducing *continuation KD*. This method starts with an easy-to-train objective function that becomes increasingly complex as the training progresses, allowing the student model to learn and gradually improve its performance.

Several works have already used KD to obtain lightweight models suitable for UAVs. For example, Li et al. [16] have applied this technique for video saliency estimation, while Liu et al. [17], Yu [18], Ding et al. [19], and Luo et al. [20] used it for object detection, object recognition, action recognition, and UAV delivery, respectively. However, to our knowledge, no work has investigated knowledge distillation to produce efficient and accurate models tailored for UAVs in the context of weed mapping.

III. Materials

This section will discuss the dataset and the preprocessing and augmentation techniques implemented.

A. Dataset

Discrimination between weeds and crop plants is a significant challenge in agricultural imaging. To address this problem, the present study relied on the WeedMap dataset proposed by Sa et al. [8], which consists of orthomosaic maps of sugar beet fields (variety Beta vulgaris "Samuela") with three classes: background, crop, and weeds. Despite the limited number of classes, the dataset demonstrates a level of complexity comparable to larger semantic segmentation datasets, such as Cityscapes [21]. This complexity stems from the subtle differences between crop and weed classes and the limited number of examples. In particular, cultivated plants occupy 15-20 pixels, while individual weeds occupy only 5-10 pixels. Therefore, using pre-trained models or additional techniques, such as data augmentation, is necessary to improve the performance of the segmentation model.

More specifically, the dataset used in this study includes eight orthomosaic maps divided into two subsets based on the location of the fields: Rheinbach in Germany and Eschikon

in Switzerland. The orthomosaic maps were further divided into tiles, resulting in 971 tiles for the Rheinbach subset and 700 tiles for the Eschikon subset. Data were acquired using two unmanned aerial vehicles: a DJI Inspire2 equipped with a RedEdge-M camera for the Rheinbach subset and a DJI Mavic Pro with a Sequoia camera for the Eschikon subset. The RedEdge-M camera acquired five channels of raw image data, including red, green, blue, near-infrared (NIR), and red edge (RE). On the other hand, the Sequoia camera acquired the same channels except for the blue channel.

B. Data Preprocessing

Although the dataset has already been thoroughly processed by the authors [8], further preprocessing is necessary. First, since the orthomosaic maps are not rectangles, they have some black areas at the edges, which generate many completely black tiles. As a first preprocessing step, these tiles were removed, reducing the dataset to 557 tiles for the Rheinbach subset and 561 for the Sequoia subset. In addition, the height of each tile of 360 is quite problematic because it must be divisible by $2^i, i > 3$ as some convolutional filters would require. For this reason, four crops of size 256×256 were extracted from each image. This also reduces the computational load.

C. Data Augmentation

The authors of the dataset implemented a *random horizontal flip* during their experiments, a commonly used augmentation technique because it does not distort the image. However, a *vertical flip* can also be used without problems for images acquired from a nadir direction. Similarly, random rotations can be applied to images with a degree range of 0 to 360. To resolve the class imbalance, *selective random rotation* was used, applying the increment only to examples containing at least one pixel of the minority class (i.e., weeds). This approach helped to increase the number of images containing weeds, improving the model's ability to learn the minority class. This technique was applied only to the Eschikon subset, which had a weed representation of only 0.166% and later increased to 0.499%. The Rheinbach subset, on the other hand, already had a weed representation of 0.706%.

IV. Methods

Two different architectures, both Transformers, were used. The first, used as a teacher, is HRNet+OCR+PSA [22], [23]. The second is a modified version of Lawin [24], which is a Vision Transformer (ViT) [25] suitable for semantic segmentation. In particular, we lightened the architecture by obtaining an extra-light model, which we named "Lawin-L0". Both architectures have achieved state-of-the-art results on the Cityscapes, ADE20K, and COCO-Stuff reference datasets. However, they cannot be directly applied as-is. Weed mapping, like other precision agriculture tasks, benefits from some bands of the non-visible spectrum, particularly the NIR and RE bands. This hinders the application of deep learning models, which are typically suited to be fed RGB images. A concatenation layer with a modified first convolution layer is needed to handle other channels besides RGB.

A. Teacher

The network used as a teacher in this paper is a modified version of the HRNet+OCR+PSA architecture [23]. It comprises the HRNet+PSA backbone, a version of HRNetV2 [26] with the Polarized Self Attention (PSA) block as the attention block. High Resolution Net V2 (HRNetV2) is an ad-hoc architecture for semantic segmentation, derived from HRNet. It consists of four stages, each of which produces high-resolution features. The stages consist of repeated multi-resolution blocks. Each block consists of a multi-resolution group convolution and a multi-resolution convolution. The multi-resolution group convolution is an extension of the group convolution. It separates the input channels into multiple subsets of channels and applies a standard convolution to each subset at different spatial resolutions. In a multi-resolution convolution, on the other hand, the input and output subsets are fully connected, and each connection is a standard convolution. The output channels for each subset are the sum of the results of the convolutions on each subset of input channels.

The HRNet features are then transmitted to the decoder, which serves as the OCR (Object Contextual Representation) module [27]. The central concept of the OCR module is that the label assigned to each pixel must match the label of the object containing it. To achieve this goal, the OCR module first extracts soft object regions from feature maps and then, using an attention mechanism, computes representations of the object regions together with the pixel representations. These representations are used to improve the final representations employed to predict the segmentation map.

Starting from the basic architecture, we modified it specifically to solve the weed mapping problem (see Fig. 1). In particular, to handle additional input channels, we modified the first input layer so that it can accept not only visible channels but also non-visible channels.

B. Student

Like the teacher and other semantic segmentation models, Lawin includes an encoder and a decoder. The encoder is a type of architecture called Mix Transformer (MiT) [28], explicitly designed for semantic segmentation as an alternative to the original ViT. MiT can produce multilevel features with different resolutions, similar to CNNs, and outputs a feature map for each Transformer block. This hierarchical representation provides high-level coarse-grained and low-level fine-grained features that generally improve performance in semantic segmentation. For example, starting from an RGB image of size $3 \times H \times W$, the first Transformer block generates a feature map of size $C_1 \times \frac{H}{4} \times \frac{W}{4}$, where C_1 is the chosen embedding dimension. Then, each subsequent transformer block takes as input the feature maps of the previous block and produces a feature map $F_i \in \mathbb{R}^{C_i \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}}$, where i is the index of the block.

The decoder uses a technique called Large Window Attention Spatial Pyramid Pooling (LawinASPP), which consists of five different branches, including a pooling layer, a shortcut connection, and three large window attentions with different

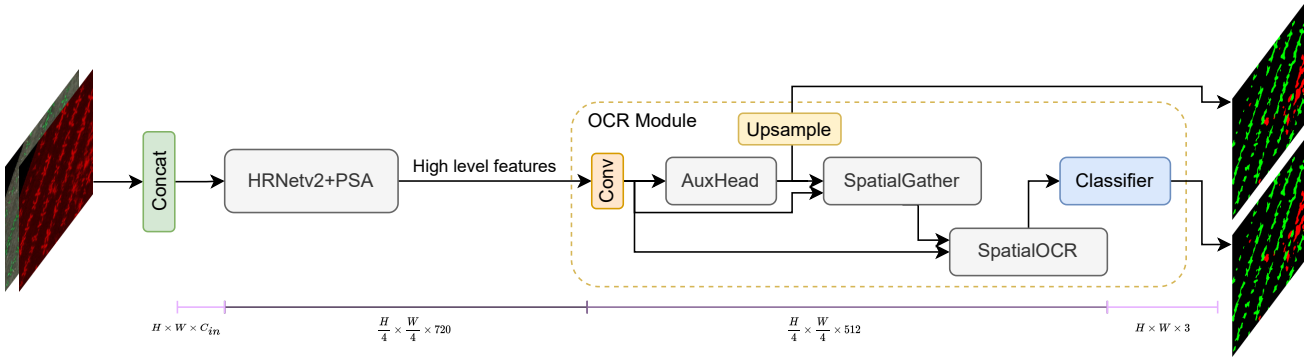

Fig. 1. HRNet+OCR+PSA modified for weed mapping ("Conv" is a convolutional layer with a 3×3 kernel, while H, W, and C_{in} are the height, width, and channels of the input images, respectively). Unlike the original version, the model accepts both visible and non-visible channels as input.

context sizes. The pooling branch handles the global context, while the three window attentions serve as local context extractors. The last two outputs of the encoder are processed with a standard multilayer perceptron and an upsampling operation. However, the first output is not processed by LawinASPP but concatenated with its output. A final linear transformation is applied to create the final segmentation map, followed by an upsampling operation. The resulting map is a probability distribution that assigns each pixel to a specific class.

As for HRNet+OCR+PSA, starting from the basic architecture of Lawin, we modified it to accept visible and non-visible channels. Moreover, to further improve performance, we propose a lighter variant of Lawin, Lawin-L0, which uses SegFormer as the encoder [28]. SegFormer, in turn, has five variants based on embedding size and model depth (B0, B1, B2, B3, B4). Lawin-L0 has a halved embedding size and a halved number of blocks in each phase compared with Lawin-B0. In addition, Lawin-B0 repeats each of the four stages twice, while Lawin-L0 repeats them only once. The embedding size in Lawin-B0 is (32, 64, 160, 256), while in Lawin-L0 it is (16, 32, 80, 128). The decoder also reflects these sizes, further reducing the computational cost. Lawin-L0 is shown in Fig. 2.

C. Vanilla Knowledge Distillation

In precision agriculture using drones, obtaining lightweight models is critical. Knowledge distillation can help achieve higher accuracy on lighter models. In particular, for weed mapping, we want to show that lightweight models can achieve comparable performance to large models when adequately trained.

In *vanilla KD*, the loss is a weighted sum of a task loss and a distillation loss:

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{task} + \alpha\mathcal{L}_{KD}$$

where $\mathcal{L}_{task}$ is the task-specific loss function for the student model, $\mathcal{L}_{KD}$ is the distillation loss, and α is a hyperparameter controlling the relative weighting between the two losses.

D. Teacher Assistant Knowledge Distillation (TAKD)

This variant of KD consists of two distillation stages [13], where the first stage involves the teacher model distilling its knowledge to an intermediate assistant model, and the second stage involves the assistant model further distilling knowledge to the final student model. This approach is designed to leverage the expertise of the teacher model while mitigating the impact of the capability gap between the teacher and student models. Introducing the assistant model as an intermediary aims to minimize the loss of information and enable a more effective transfer of knowledge to the student model. As an assistant, we used Lawin, specifically the B0 variant, which falls between the teacher and student models in terms of complexity.

E. Annealing Knowledge Distillation

This technique tries to solve the capacity gap problem by modifying the KD loss and introducing a dynamic temperature function to make the student's training gradual and smooth [14]. The process is divided into two phases: Stage I, gradual training of the student to imitate the teacher using the annealing KD loss; Stage II, fine-tuning the student with hard labels using the task loss. The resulting loss can be defined as:

$$\mathcal{L} = \begin{cases} \mathcal{L}_{KD}^{annealing}(i), & \text{Stage I: } 1 \leq \mathcal{T}_i \leq \tau_{max} \\ \mathcal{L}_{task}, & \text{Stage II: } \mathcal{T}_n = 1 \end{cases}$$

where i denotes the epoch index in the training process with n maximum epochs for Stage I and $\mathcal{T}_i$ the corresponding temperature value. At epoch (i), $\mathcal{L}_{KD}^{annealing}(i)$ is defined as:

$$\mathcal{L}_{KD}^{annealing}(i) = ||z_s(x) - \Phi(\mathcal{T}_i)z_t(x)||_2^2$$

$$\Phi(\mathcal{T}) = 1 - \frac{\mathcal{T}-1}{\tau_{max}}, 1 \leq \mathcal{T} \leq \tau_{max}, \mathcal{T} \in \mathbb{N}$$

In this case, the distillation loss is a mean squared error (MSE) between the student's logits $(z_s(x))$ and an annealed version of the teacher's logits $(z_t(x))$, obtained by multiplying them by the annealing function Φ. The annealing function is a monotonically decreasing function $\Phi : [1, \tau_{max}] \in \mathbb{N} \rightarrow [0, 1] \in \mathbb{R}$. τ_{max} represents the hyperparameter for the maximum temperature.

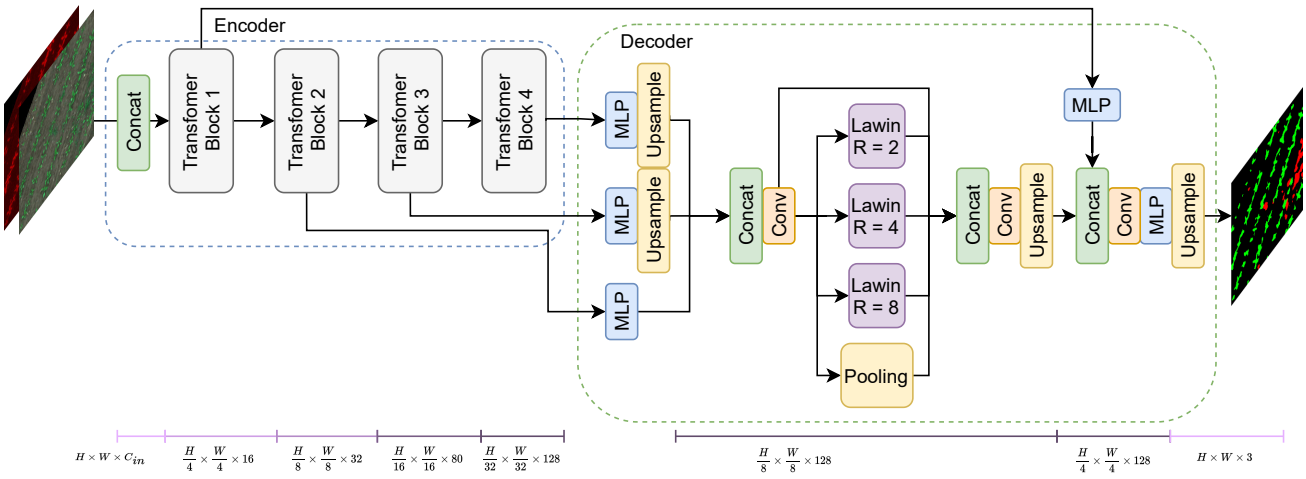

Fig. 2. Lawin-L0 ("Conv" is a convolutional layer with a 3×3 kernel, "MLP" stands for fully-connected layer, and H, W, and C_{in} are the height, width, and channels of the input images, respectively). As before, unlike the original version, the model accepts both visible and non-visible channels as input and has a reduced number of hidden channels.

F. Continuation Knowledge Distillation

This technique is based on continuation optimization, a method for solving optimization problems by gradually increasing the complexity of the objective function [15]. The idea is to start with an easy-to-train objective function that becomes increasingly complex as the training progresses, allowing the student model to learn and gradually improve its performance. The loss function is defined as:

$$\mathcal{L} = \psi(i)\mathcal{L}_{task} + (1 - \psi(i))\mathcal{L}_{KD}$$

where $\psi(i)$ is a monotonically increasing linear function $\psi : \mathbb{N} \to [0, 1] \in \mathbb{R}$. The ψ function is defined as:

$$\psi(i) = \begin{cases} \frac{i}{N_{epochs}} & \text{if } i \leq N_{epochs} \\ 1 & \text{if } i > N_{epochs} \end{cases}$$

where i is the epoch index and N_{epochs} is the number of epochs the student model will learn from the teacher.

$\mathcal{L}_{KD}$ is the distillation loss, defined as the MSE between the student's logits ($z_s(x)$) and the annealed teacher's logits ($z_t(x)$) similarly to annealing KD, but with a defined margin m:

$$\mathcal{L}_{KD} = \max\{0, ||z_s(x) - \Phi(\mathcal{T}_i)z_t(x)||_2^2 - \Phi(\mathcal{T}_i)m\}$$

where $\Phi(\mathcal{T}_i)$ is the annealing function.

V. Experiments

This section presents our experimental setup, followed by the quantitative and qualitative results of crop and weed segmentation.

A. Experimental Setup

For the Rheinbach subset, we used the same train-test subdivision applied in [8] and [29], namely [000, 001, 002, 004]–[003]. Due to the limited number of images containing weeds in the Eschikon subset test set, we opted for a different split to ensure more reliable results, i.e., [005, 007]–[006]. All channels provided in the two subsets are fed to the models. We used Adam as an optimizer for model training, with batch size 6, a maximum number of epochs of 500, and an early stop with patience 25. Specifically, the validation sets were randomly extracted from the training sets for early stopping. We used the regional mutual information [30] as the task loss, weighted by the frequency of pixel classes, as done in [8]:

$$\mathcal{L}_{RMI} = \lambda w_c \mathcal{L}_{ce}(y, p) + (1 - \lambda)\frac{1}{B}\sum_{b=1}^{B}\sum_{c=1}^{C}(-I_l^{b,c}(\mathbf{Y}; \mathbf{P}))$$

where $\lambda \in [0, 1]$ is a weight factor, $\mathcal{L}_{ce}$ is the cross-entropy, B denotes the batch size, C the number of classes, $I_l^{b,c}(\mathbf{Y}; \mathbf{P})$ is the mutual information between the ground truth and the prediction, and $\mathbf{Y}$ and $\mathbf{P}$ are the ground truth and the prediction, respectively. w_c are the class weights, calculated as:

$$w_c = \frac{FoA(c)}{\widetilde{FoA(c)}}$$

$$FoA(c) = \frac{I_c}{I}$$

where $f(x)_c$ is the probability of the true class c predicted by the model, $\widetilde{FoA(c)}$ is the median of $FoA(c)$ by varying c, I_c is the number of pixels in c, and I is the total number of pixels. The eventual application of these weights is a hyperparameter in the experiments.

In addition, we used Kullback-Leibler divergence and MSE for vanilla KD as the distillation loss, with $\alpha = 0.8$. The same hyperparameters were used for TAKD. As for annealing KD, we used an initial temperature of 0.9. In addition, it is not possible to use early stopping because the temperature is a function of epochs, so we set 50, 100, and 150 as the maximum epochs. For continuation KD, we used the same hyperparameters as for annealing KD, but given the way ψ

TABLE I
COMPARISON OF DIFFERENT KD TECHNIQUES WITH THE TEACHER AND THE UNDISTILLED MODEL FOR BOTH SUBSETS OF DATA

	Rheinbach				Eschikon			
	F1	**F1 Background**	**F1 Crop**	**F1 Weed**	**F1**	**F1 Background**	**F1 Crop**	**F1 Weed**
Teacher	0.868	0.990	0.877	0.737	0.656	0.995	0.816	0.155
No distillation	0.843	0.989	0.847	0.696	0.565	0.995	0.653	0.046
Vanilla KD	0.855	0.989	0.855	0.719	0.624	0.990	0.668	0.215
TAKD	0.850	0.988	0.842	0.720	**0.631**	0.992	0.763	0.138
Annealing KD	0.853	0.989	0.849	0.722	0.581	0.994	0.691	0.059
Continuation KD	**0.863**	0.990	0.862	0.736	0.553	0.987	0.666	0.005

is defined, the early stopping technique can be used. The experiments were performed on an RTX 3080 Ti with 12 GB of VRAM.

We used the F1 score for each class and macro-averaged to assess the models quantitatively. It was calculated as follows:

$$F1 = \frac{2TP}{2TP + FP + FN}$$

where TP stands for true positives, FP for false positives, and FN for false negatives.

B. Quantitative Results

The study's results are presented in Table I, which compares the performance of the teacher model used alone, the student model without teacher knowledge distillation, and the student model with knowledge distillation. We do not show the results obtained with all possible hyperparameter configurations for better readability, but only the best ones obtained. Furthermore, it is worth noting that the different KD methods did not modify the number of parameters of the resulting models but only the training procedure.

The study found that in both subsets, the use of knowledge distillation improved performance. Interestingly, the vanilla KD approach was sufficient to improve the score. Continuation KD outperformed the other models for the Rheinbach subset, with an average F1 score of only 0.005 lower than the teacher model and 0.001 lower for the weed class. On the other hand, the Eschikon subset presented difficulties due to the differences between the crops of the three collected fields, highlighting the difficulty of generalization in weed mapping. However, the teacher assistant technique showed promise, improving performance by up to 7%. The F1 score for weeds also showed a substantial increase, from 0.046 without KD to 0.138 with TAKD. In addition, the F1 score for crop class showed a significant increase from 0.653 to 0.753. Although there was a reduction in the background class score, the score was still high enough to make the reduction insignificant. State-of-the-art models for the WeedMap dataset include DeepLabV3 [11], which obtained an F1 score of 0.81 on the Rheinbach subset. Instead, while not using F1 in their experiments [8], replicating SegNet scores 0.445 on Eschikon and 0.836 on Rheinbach. Therefore, our lightweight model outperforms even the dataset's state-of-the-art.

Regarding computational time and complexity, Lawin-L0 has a relatively low amount of computational operations, measured in GMacs (giga multiply-accumulated operations), equal to 0.5 GMacs, and a relatively low number of parameters, measured in millions, equal to 0.98 million parameters. On the other hand, HRNet+OCR+PSA has significantly higher computational requirements, with 42.04 GMacs and 75.74 million parameters. Despite the substantial reduction in parameters, a satisfactory level of accuracy can be achieved.

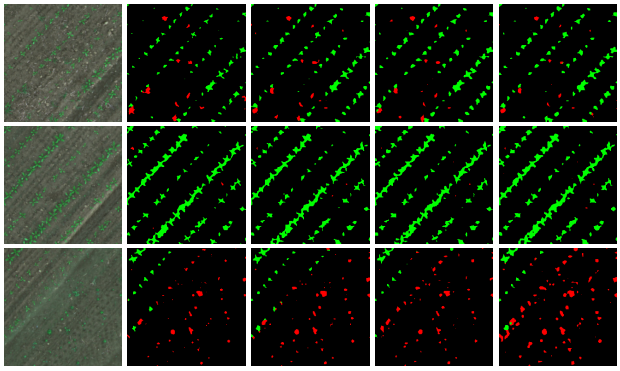

Fig. 3. Segmentation examples performed on test field [003] (black = background, green = crop, red = weed).

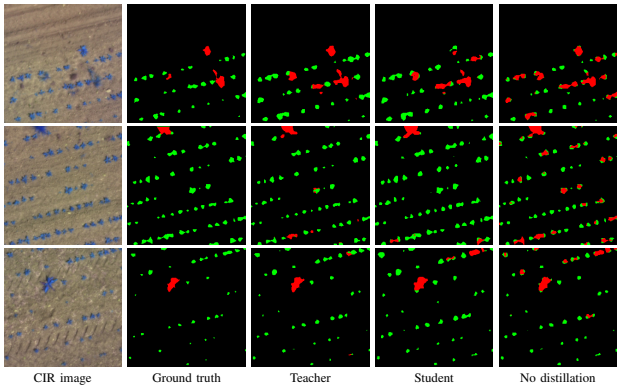

Fig. 4. Segmentation examples performed on test field [006] (black = background, green = crop, red = weed). CIR stands for color infrared.

C. Qualitative Evaluation

A qualitative assessment of weed mapping found that the segmentation maps obtained from the students' model were the same quality as those obtained from the teacher model. This is reflected in the F1 score obtained from both models. This indicates that the students' model can learn effectively from the teacher model and produce accurate weed mapping results. Examples of segmentation maps obtained as output

from the best execution of student Lawin-L0 and teacher HRNet+OCR+PSA on Rheinbach are shown in Fig. 3. In the Rheinbach subset, the segmentation maps reveal no apparent visual difference in the F1 score, despite all models performing well. However, distinct differences are observed for the weed class in the Eschikon subset, shown in Fig. 4. The models show an imbalance toward the weed class, with high recall, low precision, and many false positives. In particular, this phenomenon is more pronounced in the undistilled model than in the distilled model. The segmentation maps produced by the distilled model resemble those of the teacher model, indicating the significant influence of the teacher model on students' predictions.

VI. Conclusion

Our study demonstrated that knowledge distillation in the context of drone-based weed mapping could be effectively used to train an extremely lightweight model with only 0.5 GMacs. Our results indicate that this model can provide high-level performance while maintaining a short inference time. This makes them ideal for mobile platforms such as drones or ground control stations, which can also be smartphone devices. In particular, we have shown that the student model can learn from the teacher model and produce accurate results. Applying knowledge distillation to the Rheinbach subset resulted in a relatively modest 2% increase in the F1 score. However, the technique proved more effective for the more challenging Eschikon subset, where a significant 7% improvement was achieved. This highlights the practical value of knowledge distillation in this particular context. A potential future direction of this research could be to apply knowledge distillation to other tasks similar to weed mapping. This would allow us to evaluate the effectiveness of our approach further and explore its potential applications in a broader range of contexts where lightweight models are critical (for example, in crowd flow detection [31]).

In conclusion, developing effective and efficient computer vision algorithms on drones can significantly improve weed management practices, leading to more sustainable and efficient farming practices. By enabling farmers to quickly and easily identify infested areas and prioritize control efforts, this technology has significant implications for precision agriculture, ultimately increasing profitability and environmental sustainability.

Acknowledgment

The research of Pasquale De Marinis is funded by a Ph.D. fellowship within the framework of the Italian "D.M. n. 352, April 9, 2022" - under the National Recovery and Resilience Plan, Mission 4, Component 2, Investment 3.3 - Ph.D. Project "Computer Vision techniques for sustainable AI applications using drones", co-supported by "Exprivia S.p.A." (CUP H91I22000410007).

References

[1] FAO, "How to Feed the World in 2050. Insights from an Expert Meet," *FAO*, 2009.

[2] S. G. Vougioukas, "Agricultural Robotics," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 2, pp. 365–392, 2019.

[3] K. Danilchenko and M. Segal, "An efficient connected swarm deployment via deep learning," in *Proceedings of the 16th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 25. IEEE, 2021, p. 1–7. [Online]. Available: http://dx.doi.org/10.15439/2021F001

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

[5] I. Sa, Z. Chen, M. Popović, R. Khanna, F. Liebisch, J. Nieto, and R. Siegwart, "Weednet: Dense Semantic Weed Classification Using Multispectral Images and Mav for Smart Farming," *IEEE robotics and automation letters*, vol. 3, no. 1, pp. 588–595, 2017.

[6] B. Hobba, S. Akıncı, and A. H. Göktogan, "Efficient Herbicide Spray Pattern Generation for Site-Specific Weed Management Practices Using Semantic Segmentation on UAV Imagery," in *Australasian Conference on Robotics and Automation (ACRA-2021)*, 2021, pp. 1–10.

[7] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge Distillation: A Survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, Jun. 2021.

[8] I. Sa, M. Popović, R. Khanna, Z. Chen, P. Lottes, F. Liebisch, J. Nieto, C. Stachniss, A. Walter, and R. Siegwart, "WeedMap: A Large-Scale Semantic Weed Mapping Framework Using Aerial Multispectral Imaging and Deep Neural Network for Precision Farming," *Remote Sensing*, vol. 10, no. 9, p. 1423, 2018.

[9] P. Lottes, J. Behley, N. Chebrolu, A. Milioto, and C. Stachniss, "Joint Stem Detection and Crop-Weed Classification for Plant-Specific Treatment in Precision Farming," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 8233–8238.

[10] M. Á. Chicchón Apaza, H. M. B. Monzón, and R. Alcarria, "Semantic Segmentation of Weeds and Crops in Multispectral Images by Using a Convolutional Neural Networks Based on U-Net," in *International Conference on Applied Technologies*. Springer, 2019, pp. 473–485.

[11] W. Ramirez, P. Achanccaray, LF. Mendoza, and MAC. Pacheco, "Deep Convolutional Neural Networks for Weed Detection in Agricultural Crops Using Optical Aerial Images," in *2020 IEEE Latin American GRSS & ISPRS Remote Sensing Conference (LAGIRS)*. IEEE, 2020, pp. 133–137.

[12] S. I. Moazzam, U. S. Khan, W. S. Qureshi, M. I. Tiwana, N. Rashid, W. S. Alasmary, J. Iqbal, and A. Hamza, "A Patch-Image Based Classification Approach for Detection of Weeds in Sugar Beet Crop," *IEEE access : practical innovations, open solutions*, vol. 9, pp. 121 698–121 715, 2021.

[13] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved Knowledge Distillation via Teacher Assistant," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 5191–5198, Apr. 2020.

[14] A. Jafari, M. Rezagholizadeh, P. Sharma, and A. Ghodsi, "Annealing Knowledge Distillation," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr. 2021, pp. 2493–2504.

[15] A. Jafari, I. Kobyzev, M. Rezagholizadeh, P. Poupart, and A. Ghodsi, "Continuation KD: Improved Knowledge Distillation through the Lens of Continuation Optimization," Dec. 2022.

[16] J. Li, K. Fu, S. Zhao, and S. Ge, "Spatiotemporal Knowledge Distillation for Efficient Estimation of Aerial Video Saliency," *IEEE Transactions on Image Processing*, vol. 29, pp. 1902–1914, 2020.

[17] B.-Y. Liu, H.-X. Chen, Z. Huang, X. Liu, and Y.-Z. Yang, "ZoomInNet: A Novel Small Object Detector in Drone Images with Cross-Scale Knowledge Distillation," *Remote Sensing*, vol. 13, no. 6, p. 1198, Jan. 2021.

[18] G. Yu, "Data-Free Knowledge Distillation for Privacy-Preserving Efficient UAV Networks," in *2022 6th International Conference on Robotics and Automation Sciences (ICRAS)*, Jun. 2022, pp. 52–56.

[19] M. Ding, N. Li, Z. Song, R. Zhang, X. Zhang, and H. Zhou, "A Lightweight Action Recognition Method for Unmanned-Aerial-Vehicle Video," in *2020 IEEE 3rd International Conference on Electronics and Communication Engineering (ICECE)*, Dec. 2020, pp. 181–185.

[20] "KeepEdge: A Knowledge Distillation Empowered Edge Intelligence Framework for Visual Assisted Positioning in UAV Delivery," https://ieeexplore.ieee.org/abstract/document/9732222/.

[21] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.

[22] A. Tao, K. Sapra, and B. Catanzaro, "Hierarchical Multi-Scale Attention for Semantic Segmentation," May 2020.

[23] H. Liu, F. Liu, X. Fan, and D. Huang, "Polarized Self-Attention: Towards High-quality Pixel-wise Regression," Jul. 2021.

[24] H. Yan, C. Zhang, and M. Wu, "Lawin Transformer: Improving Semantic Segmentation Transformer with Multi-Scale Representations via Large Window Attention," *arXiv preprint arXiv:2201.01615*, 2022.

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:2010.11929*, 2020.

[26] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, "High-Resolution Representations for Labeling Pixels and Regions," Apr. 2019.

[27] Y. Yuan, X. Chen, and J. Wang, "Object-Contextual Representations for Semantic Segmentation," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, vol. 12351, pp. 173–190.

[28] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[29] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected Crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[30] S. Zhao, Y. Wang, Z. Yang, and D. Cai, "Region Mutual Information Loss for Semantic Segmentation," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[31] G. Castellano, E. Cotardo, C. Mencar, and G. Vessio, "Density-Based Clustering with Fully-Convolutional Networks for Crowd Flow Detection from Drones," *Neurocomputing*, 2023.