POLITECNICO DI TORINO
Repository ISTITUZIONALE

Domain Generalization for Crop Segmentation with Standardized Ensemble Knowledge Distillation

Original
Domain Generalization for Crop Segmentation with Standardized Ensemble Knowledge Distillation / Angarano, S., Martini, M., Navone, A., Chiaberge, M.. - ELETTRONICO. - (2024), pp. 5450-5459. (Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops Seattle, WA (USA) 17-18 June 2024) [10.1109/CVPRW63382.2024.00554].

Availability:
This version is available at: 11583/2991002 since: 2024-10-01T13:03:33Z

Publisher:
IEEE

Published
DOI:10.1109/CVPRW63382.2024.00554

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
IEEE postprint/Author's Accepted Manuscript

©2024 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

12 October 2026

This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Domain Generalization for Crop Segmentation with Standardized Ensemble Knowledge Distillation

Simone Angarano Mauro Martini Alessandro Navone
Marcello Chiaberge
Politecnico di Torino
Corso Duca degli Abruzzi, 24, 10129, Turin, Italy
name.surname@polito.it

Abstract

In recent years, precision agriculture has gradually oriented farming closer to automation processes to support all the activities related to field management. Service robotics plays a predominant role in this evolution by deploying autonomous agents that can navigate fields while performing tasks such as monitoring, spraying, and harvesting without human intervention. To execute these precise actions, mobile robots need a real-time perception system that understands their surroundings and identifies their targets in the wild. Existing methods, however, often fall short in generalizing to new crops and environmental conditions. This limit is critical for practical applications where labeled samples are rarely available. In this paper, we investigate the problem of crop segmentation and propose a novel approach to enhance domain generalization using knowledge distillation. In the proposed framework, we transfer knowledge from a standardized ensemble of models individually trained on source domains to a student model that can adapt to unseen realistic scenarios. To support the proposed method, we present a synthetic multi-domain dataset for crop segmentation containing plants of variegate species and covering different terrain styles, weather conditions, and light scenarios for more than 70,000 samples. We demonstrate significant improvements in performance over state-of-the-art methods and superior sim-to-real generalization. Our approach provides a promising solution for domain generalization in crop segmentation and has the potential to enhance a wide variety of agriculture applications.

1. Introduction

In the last two decades, scientific research in precision agriculture has significantly evolved its automatic and self-managed processes. Automation has been analyzed through four essential requirements: increasing productivity, allocating resources reasonably, adapting to climate change, and avoiding food waste [40]. Recently, deep learning solutions led to new technological trends in all these tasks, providing competitive advantages for crop monitoring and managing [32]. Autonomous robots equipped with perception systems can assist or replace human operators in agricultural tasks such as harvesting [4], spraying [11], and vegetative assessment [13], reducing human labor and enhancing operational safety. Various computer vision methods have been proposed for navigating and monitoring row crops, most of which are based on semantic segmentation [22]. Real-time crop segmentation can be used to identify objects on different scales: detailed leaf disease [26], single fruits or branches [30], crop rows [1], and entire fields [31]. It has also been exploited for autonomous navigation [1], combined with waypoint generation [33] or sensorimotor agents [25].

However, crop segmentation presents two main challenges. First, changing weather, lighting, terrain, and crop types pose a major obstacle to generalization. Supervised training methodologies usually reach remarkable results in well-defined experimental settings but struggle to yield good results where the data distribution changes [9]. However, robustness in realistic scenarios can be enhanced using frameworks like domain generalization (DG). DG is a set of representation learning techniques that aims to train models capable of generalizing to unseen domains, i.e., out-of-distribution data. Several DG methodologies have been presented in the last years, although often limiting their scope to classification on toy datasets [41]. Applying generalization methods to realistic tasks is still limited to a few attempts [7, 20, 24]. Moreover, the considered domains are often limited to stylistic changes, overlooking more radical correlation shifts [39]. For instance, in some scenarios, brown and green positively correlate with terrain and vegetation. However, other domains present brown tree trunks

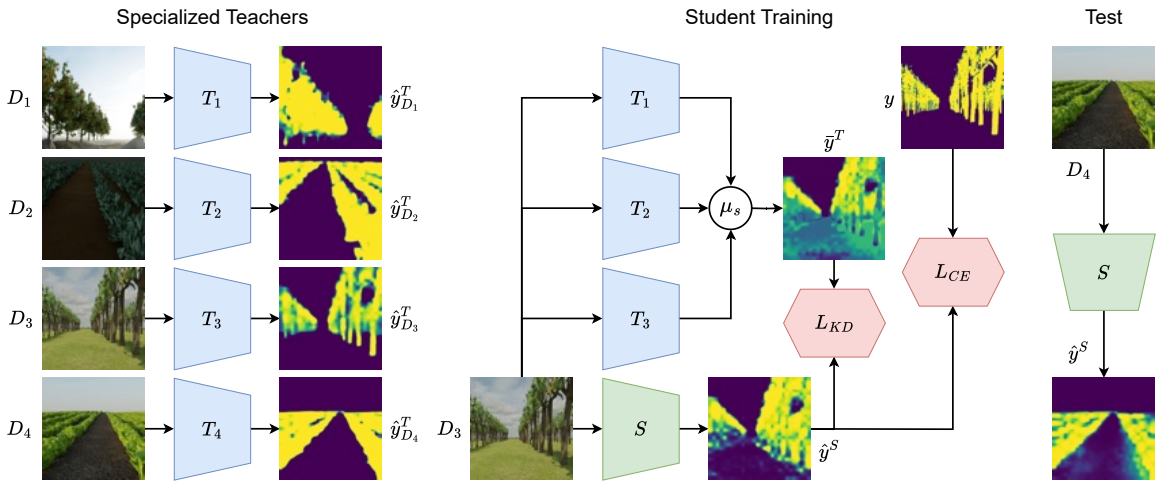

Figure 1. Schematic representation of the proposed distillation methodology for crop segmentation. Ensembled specialized teachers allow the student to obtain a standardized distillation mask ($\tilde{y}^T$) that is much more informative than the hard label (y) for robust student training. μ_s represents standardized ensembling.

and grass on the ground, inverting the correlation.

A second challenge is data availability, as no comprehensive dataset for crop segmentation across multiple scenarios is available. The reason is that on-field data collection and labeling are highly time-demanding. Hence, the only publicly available datasets focus on specific scenarios and usually include a modest number of samples. The scarcity of task-specific labeled data has recently favored the practice of synthetic data generation, leading to an additional Simulation-to-Reality (Sim2Real) gap and further compromising generalization [5].

This work aims to effectively enhance DG in crop segmentation, working towards having a single model that can generalize across different crop types and environmental conditions. It is well known that supervised neural networks exploit spurious correlations to find shortcuts in data and efficiently minimize the loss function [14]. In agricultural scenarios, these correlations can easily be found in the color of a specific species, low-level terrain textures, or background. We apply the DG framework to encourage models to learn deep, robust features without knowing the target data distribution. Moreover, recent findings have given a theoretical interpretation of the efficacy of model ensembling and knowledge distillation (KD) for robust representation learning [2]. Multiple features exist in data samples that can be used to classify them correctly, and this multi-view structure constitutes the "dark knowledge" that ensembles and KD exploit, explaining the efficacy of these methods. We investigate whether such property enhances domain and Sim2Real generalization, particularly for crop segmentation.

The proposed method distills knowledge from an ensemble of models individually trained on source domains to a student model that can adapt to unseen target domains, as depicted in Figure 1. To effectively balance the contribution of the teachers, we standardize their output logits and, in this way, avoid overconfident predictions to guide knowledge transfer. To properly validate the proposed method, we present the synthetic multi-domain dataset for crop segmentation AGRISEG, containing 11 crop types and covering different terrain styles, weather conditions, and light scenarios for more than $70,000$ samples. We conduct thorough experiments on AGRISEG and additional real-world datasets to verify the effectiveness of our method compared to other state-of-the-art solutions. The contributions of this work can be summarized as follows:

1. We propose a novel DG methodology for crop segmentation based on ensemble KD weighted by logit standardization;
2. We support our solution with AGRISEG, a rich multi-domain synthetic dataset to benchmark generalization in crop segmentation;
3. We extensively experiment on synthetic and real data to demonstrate the improvement of the proposed method on state-of-the-art solutions.

The code[1] used for the experiments and the AGRISEG dataset[2] are publicly available.

2. Related Works

Generalization to Out-of-domain (OOD) data distributions is one of the most critical requirements for real-world computer vision applications like crop segmentation. Recently,

[1] https://github.com/PIC4SeR/AgriSeg
[2] https://pic4ser.polito.it/agriseg/

Figure 2. From left to right: examples of synthetic 3D crop models used to build the AGRISEG Dataset (generic tree, zucchini, lettuce, vineyard); examples of resulting dataset images (vineyard, chard); examples of real-world test images (vineyard, miscellaneous).

rigorous validation benchmarks have been proposed to compare the advantages of different approaches and backbones for classification [3, 15]. In the meantime, segmentation across multiple scenarios has been studied, either designing massive foundation models [18] or creating new DG methods. As we aim to push the limits of generalization for efficient and easily deployable architectures, we focus on the latter approach. In particular, [28] proposed an Instance Batch Normalization (IBN) block for residual modules to avoid bias toward low-level domain-specific features like color, contrast, and texture. [27], on the same line, proposed a permuted Adaptive Instance Normalization (PAdaIN) block, which works at both low-level and high-level features, randomly swapping second-order statistics between source domains and hence regularizing the network towards invariant features. [7] proposed RobustNet, a model incorporating an Instance Selective Whitening (ISW) loss disentangling and removing the domain-specific style in feature covariance. [20] proposed to extract domain-generalized features by leveraging a variety of contents and styles using a "wild" dataset. Most recently, [37] has been the first attempt to apply KD in the DG framework for classification tasks, proposing a gradient filtering approach. [19] proposed Cross-domain Ensemble Distillation (XDED) to extract the knowledge from domain-specific teachers and obtain a general student. However, this setup was only applied to classification, while the authors used a different approach for segmentation distilling from a single teacher. Standard DG benchmarks almost solely focus on domestic environments or autonomous driving [8, 12], and generalization for crop segmentation has been addressed only in the last few years. In particular, [38] proposes a style transfer method for robust weed segmentation, considering only one crop type. [29] proposes supervised Domain Adaptation for row crop segmentation, requiring target-domain labeled data. We push the generalization concept further, including not only weather and lighting conditions but also aiming to generalize to unseen crop types without prior knowledge about the target data distribution. We take advantage of the capabilities provided by ensemble KD [2] to transfer the knowledge of domain-expert teachers to a general multi-domain student. We improve the method proposed in [19] for classification, adding logit standardization to balance the contribution of different teachers to the KD loss and applying it to real-world crop segmentation. Each teacher must be trained only once, and the method can be extended to more domains by just training a new teacher and then distilling.

3. Methodology

3.1. Domain Generalization

Given the input random variable X with values $x \in \mathcal{X}$ and the target random variable Y with values $y \in \mathcal{Y}$, the definition of *the domain* is associated with the joint probability distribution $P(X,Y)$ (P_{XY} for simplicity) over $\mathcal{X}\times\mathcal{Y}$. Supervised learning aims to train a classifier $f : \mathcal{X} \rightarrow \mathcal{Y}$ exploiting N available labeled examples of a dataset $D = (x_i, y_i)_{i=1}^{N}$ that are identically and independently distributed (i.i.d.) and sampled according to P_{XY}. The goal of the training process is to minimize the *empirical risk* associated with a loss function $l : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, +\infty)$,

$$R_{\text{emp}}(f) = \frac{1}{N}\sum_{i=1}^{N} l(f(x_i), y_i) \tag{1}$$

by learning the classifier f. Dataset D is the only available source of knowledge to learn P_{XY}. We refer to this basic learning method as empirical risk minimization (ERM) [36] and use it as a baseline for the experimentation.

In DG, a set of different K source domains $\mathcal{S} = (S_k)_{k=1}^{K}$ is used to learn a classifier f that aims at generalizing well on an unknown target domain T. Each source domain is associated with its joint probability distribution

P^k_{XY}, whereas P^S_{XY} indicates the overall source distribution learned by the classifier [41]. Indeed, DG aims to enable the classifier to predict well on out-of-distribution data, namely on the target domain distribution P^T_{XY}, by learning an overall domain-invariant distribution from the source domains seen during training.

3.2. Knowledge Distillation

KD aims at transferring the knowledge learned by a *teacher* model to a smaller or less expert *student* model. It was first proposed in [6], received greater attention after [16], and is one of the most promising techniques for model compression and regularization today. In its original formulation based on classification, KD applies an auxiliary loss to the output logits of the student $z_S(x) \in \mathbb{R}^C$, where C is the number of classes. The posterior predictive distribution of x can be formulated as:

$$P(y|x;\theta,\tau) = \frac{exp(z_y(x)/\tau)}{\sum_{i=1}^{C} exp(z_i(x)/\tau)} \tag{2}$$

where y is the label, θ is the set of parameters of the model, and τ is the temperature scaling parameter. To match the distributions of student and teacher, KD minimizes the Kullback-Leibler Divergence between the two:

$$L_{\text{KD}}(X;\theta,\tau) = \sum_{x_i \in X} \sum_{c=1}^{C} L_{\text{KD}}^{x_i,c} \tag{3}$$

$$L_{\text{KD}}^{x_i,c} = D_{\text{KL}}(P(c|x_i;\theta_T,\tau)||P(c|x_i;\theta_S,\tau)) \tag{4}$$

where X is a batch of input samples and θ_T and θ_S are the parameters of teacher and student, respectively. In this work, we apply a novel KD technique for semantic segmentation to improve models' generalization ability across domains.

3.3. Standardized Ensemble Distillation

We propose a simple yet effective training procedure based on model ensemble, KD, and logit standardization to encourage the model to learn domain-invariant features. We choose ensemble KD encouraged by the recent theory of [2] on multi-view extraction from data. Ensemble KD has been previously applied to classification in XDED [19], leveraging the separate pretraining of a teacher for each source domain and distilling the ensembled predicted logits. We aim to apply the same intuition to crop semantic segmentation, taking into account the differences between the two tasks and the additional challenges given by the agricultural setting and the Sim2Real gap. Another important challenge of this application scenario is that domain shifts are not only given by style transfer but also by the presence of completely different crop types.

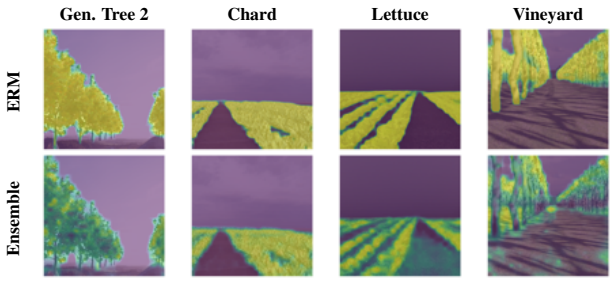

Figure 3. Comparison of ERM predictions with our ensemble of specialized teachers. While for simpler domains, the predictions of the specialized teachers agree and return a high-confidence mask, for challenging ones, the teachers give an uncertain but more informative mask that can be distilled into the student.

In our proposed method, we train a teacher for each source domain and ensemble them to create the distillation knowledge:

$$\bar{y}^T(x) = \frac{1}{D} \sum_{d=1}^{D} \hat{y}_d^T(x) \tag{5}$$

where $\hat{y}_d^T(x)$ is the predicted logits tensor for the source domain d, $\bar{y}^T$ is the ensembled teacher logits tensor, and D is the number of source domains. The motivation behind this choice is that by averaging the predictions of different specialized models, the resulting map is much more informative than the ground-truth label. As depicted in figure 3, the teacher's segmentation is less confident and often assigns non-zero probabilities to disturbing elements such as grass and background vegetation. This spurious information guides the student towards implicitly recognizing what features are more likely to be confounding at test time. This information does not overcome label supervision, as the distillation loss has a relatively low weight in the optimization process. On the contrary, if the distillation mask is very confident, the student is guided toward being more confident and implicitly incorporates the information that a certain domain is easier to segment. This effect can be tuned using a temperature factor and a weight loss.

We train the student in the standard ERM DG framework with an additional distillation loss based on the distance between the student logits and the ensembled teacher logits. To improve the effectiveness of the distillation mechanism, we design a simple mechanism to prevent the student from being biased by the overconfidence of the teachers. Indeed, each teacher is trained on a single domain and hence can fall for spurious correlations in the training data (e.g., color bias). We hence propose to standardize teacher and student logits [35] before distilling as follows:

$$\tilde{y}^T = \frac{y^T - \bar{y}^T}{\sigma(y^T) \cdot \tau}, \qquad \tilde{y}^S = \frac{\hat{y}^S - \bar{y}^S}{\sigma(\hat{y}^S) \cdot \tau} \tag{6}$$

where $\bar{y}$ is the mean, $\sigma(y)$ is the standard deviation of the

logits, and τ is the temperature. The intuition behind this choice is that optimal τ could vary across domains due to the teachers being more or less confident about their predictions. Standardization allows for adaptive calibration of logit temperature and effective domain knowledge transfer.

The distillation loss is calculated as the Kullback-Leibler divergence between teacher and student logits:

$$L_{\text{KD}}(\tilde{y}^T, \tilde{y}^S) = \frac{\tau^2}{C} \sum_{c=1}^{C} \sum_{i=1}^{W \cdot H} \phi(\tilde{y}^T_{c,i}) \cdot log\left(\frac{\phi(\tilde{y}^T_{c,i})}{\phi(\tilde{y}^S_{c,i})}\right) \quad (7)$$

where C is the number of classes and τ is the temperature.

In combination with the distillation loss, we optimize the standard cross-entropy loss between student logits and ground-truth labels y:

$$L_{\text{CE}}(y, \hat{y}^S) = -\sum_{i=1}^{C} y_i \cdot log(\hat{y}^S_i) \quad (8)$$

which for binary segmentation becomes a simple binary cross-entropy loss.

The overall loss can be written as follows:

$$L(y, \bar{y}^T, \hat{y}^S) = L_{\text{CE}}(\bar{y}, \hat{y}^S) + \lambda L_{\text{KD}}(\tilde{y}^T, \tilde{y}^S) \quad (9)$$

where λ is a weighting parameter to balance the loss components. We remark that our method adds no overhead at test time. We provide a thorough ablation of the various components of our method in 5.2 to highlight the strong improvement on previous solutions.

4. Experimental Setting

This section describes the details of the proposed synthetic AGRISEG segmentation dataset and the procedure we followed to validate the effectiveness of our DG methodology. In 4.1, we review the procedure followed to generate the AGRISEG dataset, while in section 4.2, details on the training framework and implementation are given.

4.1. Dataset

To generate the synthetic crop dataset with realistic plant textures and measurements, high-quality 3D plant models have been created using Blender[3]. A wide variety of crops have been included in the dataset to validate the segmentation performance of the model trained with the proposed DG method. Depending on the plant's height, three primary macro-categories of crops have been identified. Low crops, such as lettuce and chard, have an average height of 20-25 cm. Medium crops, such as zucchini, grow to 60 cm. Tall crops, which include vineyards and trees, can grow up to 2.5-4.5 m. Some examples of 3D plant models are shown in Figure 2.

[3] https://www.blender.org/

Crop	Samples	Type	Category↓	Height [m]
Lettuce	4800	Synthetic	Low	0.22
Chard	4800	Synthetic	Low	0.25
Lavender	5260	Synthetic	Low	0.3
Zucchini	19200	Synthetic	Medium	0.6
Cotton	4800	Synthetic	Medium	0.6
Vineyard	4800	Synthetic	Tall	1.5-2.5
Pergola Vineyard	4800	Synthetic	Tall	3.2
Apple Tree	9600	Synthetic	Tall	2.7
Pear Tree	4800	Synthetic	Tall	3.0
Generic Tree 1	4800	Synthetic	Tall	4.5
Generic Tree 2	2785	Synthetic	Tall	4.5
Vineyard [1]	500	Real	Tall	2.5
Miscellaneous	100	Real	Any	Any
VegAnn [23]	3775	Real	Any	Any

Table 1. Detailed properties for each domain of the AGRISEG dataset. The section on the top reports the synthetic crop datasets generated in simulation, while the section on the bottom the real-world ones.

Various terrains and sky models have been used to achieve realistic background and light conditions. Afterward, Blender's Python scripting functionality was used to automatically separate plants from the rest of the frame and generate a dataset of RGB images and corresponding segmentation masks. This work presents the AGRISEG dataset, composed of samples from low crops (e.g., chard and lettuce), medium crops (e.g., zucchini), and tall crops (e.g., vineyard, pear tree, and generic tall tree). Each dataset presents four sub-datasets that differ in background and terrain. Cloudy and sunny skies, diverse lighting, and shadow conditions are included. Camera position and orientation have been changed to capture diverse image samples along the whole field for each sub-dataset. Overall, the AGRISEG dataset contains more than $70,000$ samples. In the bottom rows, we also include three additional domains to validate the considered solutions on real-world data as a final test. The *Real Vineyard* dataset was originally presented in [1], but the proposed labels were coarse. Hence, we re-label the samples using the *SALT* labeling tool [4] based on Segment Anything [18]. We also add *Miscellaneous*, containing 100 samples from disparate crop types, and label it similarly. Finally, we include *VegAnn* [23], a multi-crop dataset acquired under diverse conditions for vegetation segmentation. This domain constitutes a highly different setting from the training domains, so we use it to evaluate generalization in extreme domain shifts. Details for each dataset are listed in Table 1.

4.2. Training

In this section, we report all the relevant information regarding experimental settings for model training and test-

[4] https://github.com/anuragxel/salt

Method	Gen. Tree 2	Chard	Lettuce	Vineyard	Average
Teacher	84.52 ± 0.62	95.09 ± 0.11	95.37 ± 0.10	85.51 ± 1.08	90.12 ± 0.48
ERM[36]	76.31 ± 1.53	87.63 ± 0.86	80.64 ± 5.00	67.34 ± 2.18	77.98 ± 1.62
IBN[28]	79.15 ± 1.57	88.92 ± 1.07	57.85 ± 6.00	69.11 ± 2.88	73.76 ± 2.36
ISW[7]	77.14 ± 1.77	89.44 ± 0.47	53.86 ± 8.45	68.76 ± 3.06	72.30 ± 2.30
pAdaIN[27]	75.17 ± 2.03	86.65 ± 1.75	78.05 ± 4.90	69.79 ± 1.41	77.41 ± 0.77
WildNet[20]	82.34 ± 1.55	93.68 ± 0.08	43.55 ± 6.91	72.83 ± 0.76	73.10 ± 2.70
CWD[34]	64.20 ± 8.08	84.70 ± 1.92	83.84 ± 2.79	62.88 ± 2.49	73.90 ± 2.17
WCTA[38]	75.09 ± 0.94	86.66 ± 1.93	70.73 ± 10.85	66.57 ± 3.20	74.76 ± 2.25
KDDG[37]	80.13 ± 1.61	87.67 ± 1.66	74.16 ± 3.37	65.55 ± 1.18	76.88 ± 0.68
XDED[19]	77.18 ± 2.20	88.62 ± 0.71	75.82 ± 5.12	70.48 ± 1.00	$\underline{78.03 \pm 1.85}$
Ours	78.84 ± 1.24	88.35 ± 1.27	78.04 ± 3.46	72.21 ± 1.02	$\mathbf{79.36 \pm 1.04}$

Table 2. Comparison between the proposed methodology and other state-of-the-art DG algorithms for semantic segmentation adopting the leave-one-out DG validation procedure described in 4.2. We report the Intersection-over-Union (IoU) metric (in %) for each result as mean and standard deviation. The best and second-best overall results are highlighted and underlined, respectively.

ing: data preprocessing, hyper-parameter search, and implementation. We repeat each training five times with different and randomly generated seeds to obtain statistically relevant metrics. Each of our benchmark's results is reported as mean and standard deviation.

4.2.1 Data Preprocessing

We preprocess input images through the ImageNet standard normalization [10] to use pretrained weights. We apply the same data augmentation to all the experiments, consisting of random cropping with a factor in the range $[0.5, 1]$ and flipping with a probability of 50%. We don't use random jitter, contrast, and grayscale, which are common practices in DG. We instead draw inspiration for the WCTA stylization method proposed in [38] for weed segmentation. We apply it randomly with probability $p = 0.001$ and call our version pWCTA. The reason is that we don't tackle just a shift in style but also in context (the change of crop type), and stronger stylization could over-regularize training. Experiments confirm that our choice leads to enhanced generalization on the proposed dataset.

4.2.2 Hyper-parameters

We conduct a random search to determine the optimal training hyper-parameters for the ERM DG baseline. We define a range of values for continuous arguments and a set of choices for discrete ones and select the best combination via the *training-domain validation set* strategy proposed in [15]. It consists of picking the model that maximizes the metric (in our case, IoU) on a validation split of the training set (in our case, 10%, uniform across domains) at the end of each epoch.

We choose a batch size $B = 64$ and set the number of training epochs to 50. We choose temperature $\tau = 2$ and weight $\lambda = 0.1$ for L_{KD}. We use AdamW [21] as the optimizer with a weight decay of 10^{-5}. The learning rate is scheduled with a polynomial decay between 5×10^{-5} and 5×10^{-6}. We compare to state-of-the-art methodologies running the same hyper-parameter search when tuning is necessary. We apply IBN [28] and ISW[7] to the first two blocks of the backbone, while pAdaIN [27] is applied to all the layers with a probability of 10^{-3}. The ISW loss is weighted by a factor of 10^{-2}, while XDED [19] is applied with a weight of 10^{-3}, a τ of 2, and in combination with UniStyle feature whitening.

4.2.3 Implementation

To tackle a realistic real-time application and following previous work on crop segmentation [1], we choose MobileNetV3 [17] with an LR-ASPP segmentation head[17] as the model architecture. This choice provides an optimal trade-off between performance and efficiency, exploiting effective modules such as depth-wise convolutions, channel-wise attention, and residual skip connections. We train models starting from ImageNet pretrained weights, so the input size is fixed to $(224, 224)$. The considered state-of-the-art DG methodologies are taken from the official repositories when available or reimplemented. All the training runs are performed on a single Nvidia RTX 3090 GPU.

5. Results

In this section, we present the main results of the experimentation conducted to evaluate the effectiveness of the proposed methodology. First, we compare our distillation-based approach with recent and promising DG and semantic segmentation alternatives. Inspired by popular datasets for image classification, we select four domains (*Generic Tree 2*, *Chard*, *Lettuce*, and *Vineyard*) and evaluate all the

Method	Pear Tree	Zucchini	Real Vineyard	Real Misc.	VegAnn	Average
Teacher	90.84 ± 0.30	90.42 ± 0.11	69.20 ± 2.86	54.39 ± 3.71	85.70 ± 0.88	78.11 ± 1.57
ERM[36]	82.11 ± 0.93	86.11 ± 0.15	51.51 ± 7.27	67.48 ± 0.97	61.92 ± 0.93	69.83 ± 1.33
IBN[28]	82.24 ± 0.64	86.06 ± 0.07	52.04 ± 3.98	67.97 ± 2.21	63.06 ± 1.51	70.27 ± 1.68
ISW[7]	82.31 ± 0.83	86.03 ± 0.11	55.46 ± 2.83	67.68 ± 2.04	63.29 ± 1.42	70.95 ± 1.45
pAdaIN[27]	82.67 ± 0.77	85.96 ± 0.20	53.02 ± 6.26	63.39 ± 2.00	62.56 ± 1.26	69.52 ± 2.10
WildNet[20]	88.50 ± 0.31	86.32 ± 0.10	37.30 ± 0.61	72.15 ± 0.27	42.62 ± 1.36	65.38 ± 0.53
CWD[34]	79.52 ± 0.54	85.84 ± 0.07	50.83 ± 4.37	65.27 ± 3.28	61.18 ± 1.21	68.53 ± 1.89
WCTA[38]	81.80 ± 0.82	85.87 ± 0.32	54.83 ± 1.50	64.81 ± 3.99	63.22 ± 2.17	70.11 ± 1.76
KDDG[37]	81.69 ± 0.50	86.22 ± 0.13	55.99 ± 2.61	62.60 ± 4.23	63.49 ± 0.63	70.00 ± 1.62
XDED[19]	82.04 ± 0.56	86.11 ± 0.25	56.10 ± 7.30	66.92 ± 1.93	64.09 ± 0.80	$\underline{71.05 \pm 1.41}$
Ours	83.83 ± 0.11	86.39 ± 0.04	57.21 ± 3.49	69.84 ± 1.34	65.00 ± 1.55	$\mathbf{72.45 \pm 1.31}$

Table 3. Comparison between the proposed methodology and other state-of-the-art DG algorithms on additional target domains. We train the models on all four domains chosen for the previous benchmark. We report IoU (in %) on the unseen domains as mean and standard deviation. The best and second-best results are highlighted and underlined, respectively.

methodologies by training on three domains and testing on the fourth. The domains are selected to cover different crop dimensions and visual characteristics and guarantee a challenging generalization benchmark. Then, we perform an additional evaluation by training models on all four datasets and testing on five additional target domains (*Pear Tree*, *Zucchini*, *Real Vineyard*, *Real Miscellaneous*, and *VegAnn*). In addition, we conduct a small ablation study to investigate the role of different components in our methodology and the importance of specialized single-domain teachers.

5.1. DG Benchmark

We run the leave-one-out DG benchmark described in 4.2 and report the results with their mean and standard deviation in Table 2. On average, our ensemble distillation methodology is 1.3% better than the second-best compared solution (XDED), which also distills from a set of specialized teachers. This strategy, hence, proves to give insightful information to the student and makes it less biased towards domain-specific features. The results for ERM are quite balanced across domains, proving the strong validity of this method despite its simplicity. Other DG methods, even though obtaining superior results in specific domains, are, on average, suboptimal. This failure could be due to the methods focusing on features that are extremely beneficial for some specific scenarios but useless to others. Our method, instead, retains consistently good performance in all domains thanks to the insights given by the ensembled teachers.

To further validate the generalization capability of our method, we construct a more challenging benchmark by using five unseen test domains (*Pear Tree*, *Zucchini*, *Real Vineyard*, *Real Miscellaneous* and *VegAnn*). The models are trained and validated on all four datasets used for the previous benchmark. In this test, we also investigate the Sim-to-Real gap. The results are reported in Table 3, where we also include the teachers' performance as an upper bound. Teacher IoU is lower for real datasets as they are more challenging and contain fewer samples. On average, our ensemble distillation methodology is 1.4% better than the second-best compared solution (XDED), confirming the outcome of the previous benchmark. Moreover, our method retains the best performance on all real domains. This result enforces previous considerations on the generalization ability of KD without any additional layers or computation at inference time. ERM obtains acceptable results in all domains but is surpassed by a significant margin by other methods like IBN and ISW. These results suggest ERM and state-of-the-art DG methods suffer Sim2Real more than ours. Indeed, the change from synthetic to real crops further widens the domain gap between different crops and backgrounds. Another interesting insight can be found in the standard deviations, as our method obtains one of the smallest values. WildNet performs badly on *Real Vineyard* and *VegAnn* while obtaining satisfactory results on synthetic ones. Its small standard deviation suggests that the multiple training losses applied during training could have an over-regularizing effect on the process. On the contrary, our approach finds the best trade-off between regularization and learning. In the next section, we investigate the contribution of different elements to this result.

5.2. Ablation Study

We conduct an ablation study to investigate the effect of different components on the generalization capability of our methodology. We also highlight the main differences between our approach and XDED [19] regarding methodological components and performance. In particular, we consider distillation strategy, logit standardization, and pWCTA. We also try substituting the specialized teachers with an ensemble of ERM models trained on all source

Method	Teacher	pWCTA	Logit Std	Vineyard Real	Misc. Real	VegAnn	Average
ERM	✗	✗	✗	51.51 ± 7.27	67.48 ± 0.97	61.92 ± 0.93	60.30 ± 3.06
KD	ERM	✗	✗	61.27 ± 1.03	63.70 ± 1.72	64.99 ± 0.47	63.32 ± 1.07
XDED [19]	Ens.	✗	✗	56.10 ± 7.30	66.92 ± 1.93	64.09 ± 0.80	62.37 ± 3.34
Ours	Ens.	0.001	✗	57.01 ± 2.53	69.07 ± 0.69	65.36 ± 1.12	$\underline{63.81 \pm 1.45}$
	Ens.	0.001	✓	57.21 ± 3.49	69.84 ± 1.34	65.00 ± 1.55	$\mathbf{64.02 \pm 2.13}$

Table 4. Ablation study highlighting the contribution of different design choices. We evaluate the effect of KD, domain-expert teachers, pWCTA, and logit standardization. We report IoU (in %) for each result as mean and standard deviation. The best and second-best results are highlighted and underlined, respectively.

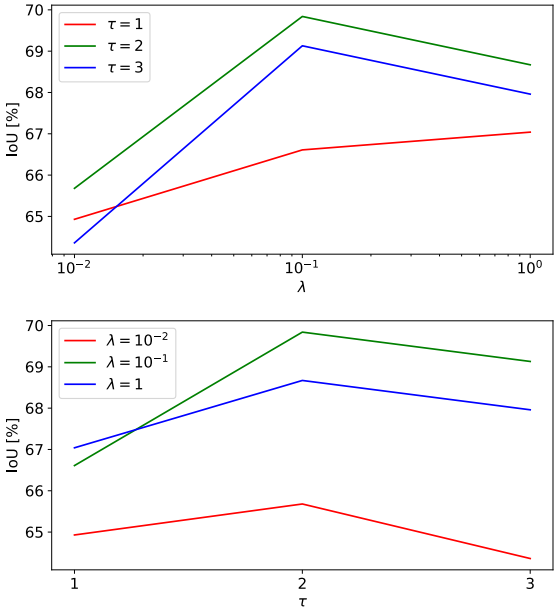

Figure 4. Ablation study on the hyper-parameters λ and τ. The reported IoU value is relative to the *Real Miscellaneous* domain and is averaged on three runs. We represent two views of the results for better readability.

domains. The results are reported in Table 4, in which we included ERM as a baseline.

First, the results confirm that applying distillation improves simple ERM without requiring additional computation at inference time. Moreover, plain ensemble distillation cannot bridge the strong Sym2Real gap. This failure is probably due to the unbalanced contribution of different teachers, which can lead to transferring domain-dependent biases to the student. To contrast this risk, we standardize distillation logits and apply pWCTA with a low probability to avoid overfitting. Our methodology outperforms ERM distillation, making the best out of specialized teachers. As depicted in Fig. 3, the distillation masks are less confident, giving the student a better understanding of what parts of the image are more likely to confound the predictor.

We further inspect the effect of the method's hyper-parameters on generalization capabilities. We vary the distillation loss weight λ and the temperature τ and report the results on the *Real Miscellaneous* domain in Fig. 4. The graphs show that our choice ($\lambda = 10^{-1}, \tau = 2$) is the optimal balance that ensures regularization without constraining the student. As reported in our benchmarks, this yields good generalization across various synthetic and real domains.

6. Conclusions

In this work, we proposed a novel method to tackle the problem of DG for crop semantic segmentation in realistic scenarios. We demonstrated that our distillation method represents a competitive approach for transferring domain-specific knowledge learned from multiple teacher models to a single student without any overhead at inference time. Moreover, we proposed logit standardization to adapt ensembled knowledge to the student, balancing overconfident predictions and penalizing spurious correlations. Each teacher must be trained only once, and the method can be extended to more domains by just training a new teacher and then distilling. Extensive experimentation has been conducted on the novel multi-crop synthetic dataset AGRISEG and on real test data to demonstrate the overall generalization boost provided by our training method. Moreover, we conducted an ablation study to highlight the role of different components in our solution. The superior results provided by our method show how pairing ensembled KD and DG can lead to robust perception models for realistic tasks in precision agriculture. Future works will progressively add more domains and DG methods to the AGRISEG benchmark. We will also include more real-world labeled data to guarantee a deeper investigation of the use of synthetic data for robust generalization in agriculture.

Acknowledgements

This work has been developed with the contribution of the Politecnico di Torino Interdepartmental Centre for Service Robotics (PIC4SeR).

References

[1] Diego Aghi, Simone Cerrato, Vittorio Mazzia, and Marcello Chiaberge. Deep semantic segmentation at the edge for autonomous navigation in vineyard rows. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3421–3428. IEEE, 2021. 1, 5, 6

[2] Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *The Eleventh International Conference on Learning Representations*, 2022. 2, 3, 4

[3] Simone Angarano, Mauro Martini, Francesco Salvetti, Vittorio Mazzia, and Marcello Chiaberge. Back-to-bones: Rediscovering the role of backbones in domain generalization. *arXiv preprint arXiv:2209.01121*, 2022. 3

[4] C Wouter Bac, Eldert J van Henten, Jochen Hemming, and Yael Edan. Harvesting robots for high-value crops: State-of-the-art review and challenges ahead. *Journal of Field Robotics*, 31(6):888–911, 2014. 1

[5] Ruud Barth, Joris IJsselmuiden, Jochen Hemming, and Eldert J Van Henten. Data synthesis methods for semantic segmentation in agriculture: A capsicum annuum dataset. *Computers and electronics in agriculture*, 144:284–296, 2018. 2

[6] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006. 4

[7] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11580–11590, 2021. 1, 3, 6, 7

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 3

[9] Gabriela Csurka. *Domain adaptation in computer vision applications*. Springer, 2017. 1

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 6

[11] Deepak Deshmukh, Dilip Kumar Pratihar, Alok Kanti Deb, Hena Ray, and Nabarun Bhattacharyya. Design and development of intelligent pesticide spraying system for agricultural robot. In *Hybrid Intelligent Systems: 20th International Conference on Hybrid Intelligent Systems (HIS 2020), December 14-16, 2020*, pages 157–170. Springer, 2021. 1

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012. 3

[13] Aijing Feng, Jianfeng Zhou, Earl D Vories, Kenneth A Sudduth, and Meina Zhang. Yield estimation in cotton using uav-based multi-sensor imagery. *Biosystems Engineering*, 193:101–114, 2020. 1

[14] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 2

[15] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020. 3, 6

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 4

[17] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 6

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 3, 5

[19] Kyungmoon Lee, Sungyeon Kim, and Suha Kwak. Cross-domain ensemble distillation for domain generalization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV*, pages 1–20. Springer, 2022. 3, 4, 6, 7, 8

[20] Suhyeon Lee, Hongje Seong, Seongwon Lee, and Euntai Kim. Wildnet: Learning domain generalized semantic segmentation from the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9936–9946, 2022. 1, 3, 6, 7

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[22] Zifei Luo, Wenzhu Yang, Yunfeng Yuan, Ruru Gou, and Xiaonan Li. Semantic segmentation of agricultural images: A survey. *Information Processing in Agriculture*, 2023. 1

[23] Simon Madec, Kamran Irfan, Kaaviya Velumani, Frederic Baret, Etienne David, Gaetan Daubige, Lucas Bernigaud Samatan, Mario Serouart, Daniel Smith, Chrisbin James, et al. Vegann, vegetation annotation of multi-crop rgb images acquired under diverse conditions for segmentation. *Scientific Data*, 10(1):302, 2023. 5

[24] Mauro Martini, Vittorio Mazzia, Aleem Khaliq, and Marcello Chiaberge. Domain-adversarial training of self-attention-based networks for land cover classification using multi-temporal sentinel-2 satellite imagery. *Remote Sensing*, 13(13):2564, 2021. 1

[25] Mauro Martini, Simone Cerrato, Francesco Salvetti, Simone Angarano, and Marcello Chiaberge. Position-agnostic autonomous navigation in vineyards with deep reinforcement learning. In *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, pages 477–484. IEEE, 2022. 1

[26] Somnath Mukhopadhyay, Munti Paul, Ramen Pal, and Debashis De. Tea leaf disease detection using multi-objective image segmentation. *Multimedia Tools and Applications*, 80: 753–771, 2021. 1

[27] Oren Nuriel, Sagie Benaim, and Lior Wolf. Permuted adain: Reducing the bias towards global statistics in image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9482–9491, 2021. 3, 6, 7

[28] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479, 2018. 3, 6, 7

[29] Shivam K. Panda, Yongkyu Lee, and M. Khalid Jawed. Agronav: Autonomous navigation framework for agricultural robots and vehicles using semantic segmentation and semantic line detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 6271–6280, 2023. 3

[30] Hongxing Peng, Chao Xue, Yuanyuan Shao, Keyin Chen, Juntao Xiong, Zhihua Xie, and Liuhong Zhang. Semantic segmentation of litchi branches using deeplabv3+ model. *IEEE Access*, 8:164546–164555, 2020. 1

[31] Ehsan Raei, Ata Akbari Asanjan, Mohammad Reza Nikoo, Mojtaba Sadegh, Shokoufeh Pourshahabi, and Jan Franklin Adamowski. A deep learning image segmentation model for agricultural irrigation system classification. *Computers and Electronics in Agriculture*, 198:106977, 2022. 1

[32] Chengjuan Ren, Dae-Kyoo Kim, and Dongwon Jeong. A survey of deep learning in agriculture: techniques and their applications. *Journal of Information Processing Systems*, 16 (5):1015–1033, 2020. 1

[33] Francesco Salvetti, Simone Angarano, Mauro Martini, Simone Cerrato, and Marcello Chiaberge. Waypoint generation in row-based crops with deep learning and contrastive clustering. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part VI*, pages 203–218. Springer, 2023. 1

[34] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5311–5320, 2021. 6, 7

[35] Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. *arXiv preprint arXiv:2403.01427*, 2024. 4

[36] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999. 3, 6, 7

[37] Yufei Wang, Haoliang Li, Lap-pui Chau, and Alex C Kot. Embracing the dark knowledge: Domain generalization using regularized knowledge distillation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2595–2604, 2021. 3, 6, 7

[38] Jan Weyler, Thomas Läbe, Federico Magistri, Jens Behley, and Cyrill Stachniss. Towards domain generalization in crop and weed segmentation for precision farming robots. *IEEE robotics and automation letters*, 8(6):3310–3317, 2023. 3, 6, 7

[39] Nanyang Ye, Kaican Li, Haoyue Bai, Runpeng Yu, Lanqing Hong, Fengwei Zhou, Zhenguo Li, and Jun Zhu. Ood-bench: Quantifying and understanding two dimensions of out-of-distribution generalization. In *CVPR*, 2022. 1

[40] Zhaoyu Zhai, José Fernán Martínez, Victoria Beltran, and Néstor Lucas Martínez. Decision support systems for agriculture 4.0: Survey and challenges. *Computers and Electronics in Agriculture*, 170:105256, 2020. 1

[41] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 4