ConvNeXt based semi-supervised approach with consistency regularization for weeds classification

Farouq Benchallal, Adel Hafiane, Nicolas Ragot, Raphaël Canals

ConvNeXt based semi-supervised approach with consistency regularization for weeds classification

Farouq Benchallal[a], Adel Hafiane[a,c], Nicolas Ragot[b], Raphaël Canals[c]

[a] *INSA CVL, PRISME EA 4229, Bourges, 18022, Centre Val de Loire, France*
[b] *Université Tours, LIFAT EA 6300, Tours, 37200, Centre Val de Loire, France*
[c] *Université d'Orléans, PRISME EA 4229, Orléans, 45067, Centre Val de Loire, France*

Abstract

Weed recognition is an essential step for automatic weed control systems. Identifying weeds enables targeted control measures to be implemented, minimizing the use of chemicals and reducing the impact on the environment. Deep learning-based approaches proved to be effective for addressing various complex classification problems. However, to benefit fully from their capabilities, large amounts of labeled data are required, which represents a limitation for agricultural applications, consequence of the tedious and time-consuming process of data labeling. Conversely, unlabeled data could be acquired in large quantities, with relative ease. Hence, our aim is to develop robust and precise deep learning models, to carry-out the recognition and identification of weed species, using both types of data. To this end, we propose a method, that adopts the semi-supervised learning paradigm, to optimally combine labeled and unlabeled data. The method is based on a new deep neural networks architecture, which consists of a modernized convolutional encoder belonging to the family ConvNeXt and a thoroughly designed deep decoder network. This architecture, enables a successful integration of consistency regularization. The conducted experiments on DeepWeeds and 4-Weeds, showed that the semi-supervised models trained through our proposed method provide a stable and high classification performance, compared to other state-of-the-art deep learning models, which were affected negatively by the amount of labeled data available, and the presence of noise during inference. Furthermore, the effectiveness of the proposed method was demonstrated in comparison to other semi-supervised learning methods. The results obtained demonstrate the benefits of adopting the semi-supervised learning paradigm, especially in scenarios with very limited labeled data.

Keywords: semi-supervised learning, deep learning, consistency regularization, precision agriculture

1. Introduction

All around the world, weeds represent the highest potential yield loss to crops along with pathogens (fungi, bacteria, etc.) and animal pests (insects, rodents, nematodes, mites, birds, etc.) being both of lesser concern (Oerke, 2006; Chauhan, 2020).Weeds are highly competitive and

Email addresses: farouq.benchallal@insa-cvl.fr (Farouq Benchallal), adel.hafiane@insa-cvl.fr (Adel Hafiane), nicolas.ragot@univ-tours.fr (Nicolas Ragot), raphael.canals@univ-orleans.fr (Raphaël Canals)

adaptable to adverse conditions. these traits give them the ability to compete with crops at every stage of their growth for nutrients, moisture, light and space, reducing the quality and quantity of the final yield. Furthermore, the reproductive mechanism of weeds is far superior to that of crop plants, especially under unfavorable conditions. Hence weeds constantly invade fields to overcome crops (Radicetti and Mancinelli, 2021).Yield losses in crops due to weeds have substantial economic losses (Pimentel et al., 2005; Llewellyn et al., 2016; Gharde et al., 2018; Chauhan, 2020).

Nowadays, most agricultural weed management practices rely heavily on the use of herbicides which requires an extensive use of chemical compounds, that may lead to undesirable effects on the health of crop plants, soil and environment (Aktar et al., 2009). Moreover, the over-reliance on herbicides with similar modes of action caused the evolution of herbicide-resistant weeds (Chauhan, 2020). Currently there have been 521 unique cases of herbicide-resistant weeds reported globally (Heap, 2023). Growing concern over the excessive use of agrochemicals is driving the adoption of precision farming practices. These practices aim to improve the efficiency of the use of agrochemicals, reduce costs and energy consumption and promote environmental protection (Radoglou-Grammatikis et al., 2020), by leveraging site-specific automated application of herbicides and robotic weed removal. The development of automated targeting mechanisms for weed control relies on an accurate identification and recognition of unsown plant species. This is a challenging problem because of field conditions, such as variation in lighting and illumination, and the similarities between weeds and crop plants in terms of color, texture, shape (Hasan et al., 2021).

Weed species classification using traditional machine learning approaches is based on the thorough process of feature extraction and selection, to apply machine learning classifiers. Sabzi et al. (2020) extracted texture features, in addition to color and shape features, and five moment-invariant features, from ground-based images, then used several algorithms to select discriminative features, and performed weed classification using various ML-based classifiers including support vector machines and random forests. Conventional machine learning techniques require deep domain expertise to conduct the complex and time-consuming process of feature engineering. On the contrary deep learning approaches are characterized by their strong ability of extracting discriminative features automatically from data, through representation-learning (Lecun et al., 2015). The high learning capacity of deep learning models allows them to perform classification and prediction particularly well, given sufficient amounts of labeled data (Kamilaris and Prenafeta-Boldú, 2018). There are currently many successful and popular deep network architectures (Bah et al., 2018), which have been used for agricultural applications. Reedha et al. (2022) applied vision transformer-based networks, i.e., VIT-B-16 and VIT-B-32 (Dosovitskiy et al., 2020) for weed and crop plants classification using unmanned aerial vehicle (UAV) imagery. Ahmad et al. (2021) used convolutional neural networks (VGG16 (Simonyan and Zisserman, 2014), ResNet50 (He et al., 2015), InceptionV3 (Szegedy et al., 2015)) to identifiy weed species in the fields of soybean and corn, from ground-based imagery.

Recent effective saliency detection methods (Cong et al., 2019; Wang et al., 2022) favor deep learning-based techniques, for identifying the most important regions and salient objects in images. This is achieved by exploiting the multi-level features extracted using deep neural networks to produce saliency maps. Zhao and Wu (2019) proposed pyramid feature attention network to focus on high-level context-aware features, and on low-level spatial structural features, to capture contextual information and refine salient object details. High-level and low-level features are then both used to generate improved saliency maps. Saliency detection can enhance image classification and segmentation. Zeng et al. (2019) developed a single network that combines a segmentation

network and a saliency aggregation module to explicitly capture the connections between the two computer vision tasks. Co-saliency detection aims to detect salient regions from a set of related images (Cong et al., 2019), by exploring inter-image correspondence. Wei et al. (2017) presented a deep learning scheme based on a fully convolutional network for co-salient objects discovery. Similarly, co-segmentation seeks to segment common objects from a group of relevant images. Li et al. (2019) proposed a deep co-segmentation approach for segmenting common objects belonging to a common semantic class from a pair of images.

Nonetheless, to fully benefit from deep learning algorithms and exploit their potential, large labeled datasets are required, which is considered as a problem for agricultural applications because of the difficulties related to the labeling process. Conversely, unlabeled data could be acquired in large quantities with relative simplicity. One of the challenges to overcome in semi-supervised learning is not only to be able to achieve a good performance as the one we can reach using large enough fully annotated data with the same model but, going further, to be able to perform similarly or better than state-of-the-art supervised models on datasets with a limited amount of labeled data. One way to do that is to solve the dilemma between using an accurate model with a sufficient number of parameters, knowing that, only few labeled data will be available to train it. Our proposition is going in this direction. Henceforth, in this paper we propose a method, that employs SSL, to generate strong predictive deep learning models, using both types of data, for weed species classification. The following summarizes our main contributions :

- The development of a new deep neural networks architecture of the auto-encoder type, based on ConvNext encoder.
- Enabling training in semi-supervised way, by taking advantage of the reconstruction model with skip-connections to derive the most important and relevant features, mainly from unlabeled data, to achieve high and accurate classification performance, with limited amounts of labeled data.
- Integration of consistency regularization constraint, to facilitate and enhance simultaneous joint learning of image classification and the reconstruction from noisy inputs, in order to improve generalization and robustness.
- Extensive experiments on three public datasets, for a rigorous evaluation and assessment of the developed deep semi-supervised model, to demonstrate the effectiveness of our proposed method.

The remaining of this paper is structured as follows: The next section, presents the related work. The subsequent section is devoted to in-depth description of our proposed method. The forth section, describes at length the conducted experiments, as well as the obtained results in comparison to other state-of-the-art deep learning models. Finally, our conclusions and perspectives, regarding future work, are given in the last section.

2. Related Work

There exist a few studies in the literature, where the semi-supervised learning paradigm has been adopted to elaborate methods, for the identification and recognition of weeds. Kerdegari et al. (2019) proposed an approach based on semi-supervised generative adversarial networks (SGANs) for semantic segmentation of weed and crop, providing a pixel-wise classification, using aerial

multispectral imagery. Khan et al. (2021) developed an optimized semi-supervised GAN-based framework for crops and weeds classification, in UAV images. Shorewala et al. (2021) used a semi-supervised technique for the estimation of weed density and distribution, from ground-based images. The approach is comprised of two steps: an unsupervised binary segmentation step, that is applied to obtain vegetation masks; and a classification step for the identification of weeds and crops, by using a fine-tuned convolutional neural network. Homan and du Preez (2021) presented a two-fold approach including feature recognition and species classification based on deep semi-supervised learning for plant identification. Hu et al. (2021) introduced a method for training site-specific weed detection models using image synthesis and semi-supervised learning. Liu et al. (2023) proposed a semi-supervised method, which incorporates a mixed attention mechanism with an explicit aim of improving model's ability to capture important features for weeds detection in wheat fields.

Our proposed method falls under a category of semi-supervised learning, based on consistency regularization (Yang et al., 2021), which has as an objective, producing reliable high-performing models. From a semi-supervised perspective we were mostly inspired by the ladder network (Rasmus et al., 2015), because of the successful utilization of the consistency regularization. Shortly afterwards, the π-model and temporal ensembling (Laine and Aila, 2016) were introduced, which are consistency regularization methods, the first relies on the stochastic transformations of training samples and the second combines the stochastic transformations with an exponential moving average of the predictions. Following these two methods Tarvainen and Valpola (2017) proposed the mean teacher, which focused more on the structure of the networks for the consistency training. Consistency regularization methods suffer from confirmation bias because they rely on a single model within the semi-supervised architecture to generate predictions for the consistency regularization term. In the case where these predictions corresponding to the unlabeled samples are incorrect, over multiple iterations, it leads to negative impact of the semi-supervised learning performance. Ke et al. (2019) presented their approach called dual student, which focused also on the structure and proposed a sample stability constraint with respect to the unlabeled data to mitigate the effect of confirmation bias. In order to overcome the confirmation bias, we propose an approach with two joint models. The formulation of the semi-supervised learning criterion is based on consistency regularization. The latter depends on minimizing the distance between reconstructed output images and clean output images, to provide accurate learning signals for more focus on incorporating reliable information about data structure during the training process.

Moreover, the above-mentioned consistency-based methods were applied to general-purpose datasets, such as CIFAR-10 and CIFAR-100 (Krizhevsky, 2009), the development of our method was driven by the goal of accurately recognizing and identifiying weed species, from images procured under real-world conditions. We can also state that, the main difference between our method and the other consistency regularization-based methods resides in the design of the proposed deep encoder-decoder architecture that utilizes skip-connections along with a significant number of trainable parameters, which is over 100 million.

3. ConvNeXt based method

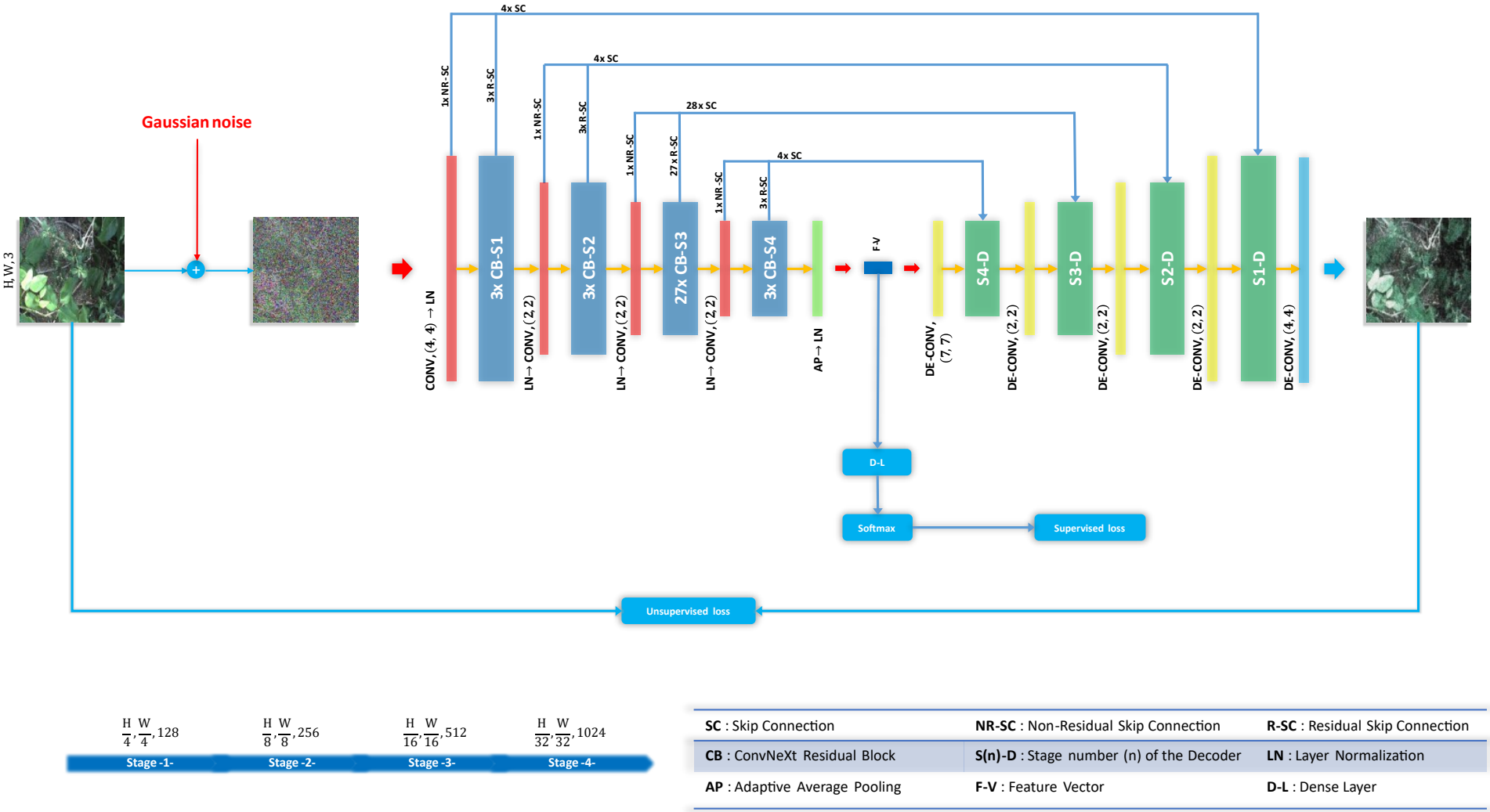
This section describes the approach used in our research work. Figure 1 gives a complete overview of the proposed deep architecture, further we present in detail the different components of this architecture. First, we describe the problem statement, then in the following subsections we present the learning strategy employed and the ConvNeXt encoding-decoding process. These

components form the foundation of our approach and ensure a thorough understanding of the developped technique.

3.1. Problem statement

In the supervised learning settings, we are given samples $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where each pair (x_i, y_i), is composed of an input x_i in $\mathcal{X}$ the space of inputs, and a label y_i in $\mathcal{Y}$ the space of outputs. In our case $\mathcal{X} \subset \mathbb{R}^d$ represents the set of images, and we have $\mathcal{Y} = \{1, \ldots, C\}$ represents the species of weeds that we want to identify, C refers to the number of classes. The join space $\mathcal{X} \times \mathcal{Y}$ is assumed to be a probabilistic space with an unknown probability measure $\mathcal{P}(x, y)$ and the data is sampled from this space, in other words $(x_i, y_i) \sim \mathcal{P}(x, y)$. The joint measure $\mathcal{P}(x, y)$ can be decomposed into a measure of the marginal distribution $\mathcal{P}(x)$, and a measure of the conditional distribution $\mathcal{P}(y|x)$. Supervised learning aims at estimating a functional relationship $x \longrightarrow y$, between a covariate $x \in \mathcal{X}$ and the class variable $y \in \{1, \ldots, C\}$, with the goal of minimizing the classification error. In the semi-supervised learning settings in addition to the labeled data $\mathcal{D}_\ell = \{(x_i, y_i)|i = 1, \ldots, n\}$ sampled from $\mathcal{P}(x, y)$, we have unlabeled data $\mathcal{D}_u = \{x_{n+j}|j = 1, \ldots, m\}$ sampled from $\mathcal{P}(x)$(Chapelle et al., 2006). When the process of acquiring labeled samples is costly and time-consuming, and obtaining unlabeled samples is inexpensive and can be done quickly, then $n \ll m$ which means that the size of the labeled portion could be much smaller than the size of the unlabeled portion. The objective of semi-supervised learning is to leverage the unlabeled data $\mathcal{D}_u$, to provide the prediction function that is characterized by its trainable parameters, with additional information about the structure of the data distribution $\mathcal{P}(x)$, which then leads to an increased performance along with better generalization to new unseen samples (Oliver et al., 2018; Yang et al., 2021). Semi-supervised learning requires that the data distribution should be under a set of assumptions. Otherwise, the prediction performance may not be improved, so that the knowledge on $\mathcal{P}(x)$, obtained through the unlabeled data carries useful information in the inference of $\mathcal{P}(y|x)$. Based on previous studies (Chapelle et al., 2006) and (Yang et al., 2021), the main assumptions associated with semi-supervised learning are as follows:

- The cluster assumption: If points are in the same cluster, they are likely to be of the same class, which means that the decision boundary should lie in a low-density region.
- The semi-supervised smoothness assumption: If two points x_1, x_2 in a high-density region are close, then so should be the corresponding outputs y_1, y_2.
- The manifold assumption: high-dimensional data lie approximately on a low-dimensional manifold. In high-dimensional spaces, the volume grows exponentially with the number of dimensions, which makes it difficult to estimate reliable densities. The second problem linked to high dimensions is that the pairwise distances tend to become more similar, and thus less expressive. If the data lie on a low-dimensional manifold, then the learning algorithms can overcome the problems related to high-dimensionality, by operating in the corresponding low-dimensional space.



SC : Skip Connection	**NR-SC** : Non-Residual Skip Connection	**R-SC** : Residual Skip Connection
CB : ConvNeXt Residual Block	**S(n)-D** : Stage number (n) of the Decoder	**LN** : Layer Normalization
AP : Adaptive Average Pooling	**F-V** : Feature Vector	**D-L** : Dense Layer

Figure 1: Overview of the approach. The proposed deep autoencoder architecture is based on the ConvNeXt-base encoder, and a specifically designed deep decoder model, with skip-connections, to enable training in semi-supervised way, by using both supervised and unsupervised losses, to ensure a more effective utilization of the labeled and unlabeled data

3.2. Semi-supervised learning process

Training in semi-supervised way is achieved through the optimization of the set of parameters related to the joint networks of our proposed architecture, by using both labeled and unlabeled data, which belong to the training subset. The optimization process is carried out, by minimizing the error of the objective function, which is composed from the sum of two learning criterions: an unsupervised loss (consistency regularization loss) and supervised loss. Our aim from the minimization of the unsupervised loss, is to utilize the characteristics of our deep joint networks, that allow for a maximum flow of information in the backward pass and thereby increasing the effect of influencing all the weights of the networks, to learn meaningful, informative, robust representations for the reconstruction of clean output images and subsequently improve the ability of generalization, to enhance the classification performance. The function we choose to represent the unsupervised term (consistency regularization term), is the L2 loss, and it takes into consideration the labeled and unlabeled portions of the training subset. Its definition is given in the following way:

$$l_{unsup}\left(x_i, \widetilde{x}_i\right) = \|x_i - \tilde{x}_i\|_2^2 = \left(x_i - \tilde{x}_i\right)^2 \tag{1}$$

Where x_i is the input image and $\widetilde{x}_i$ is the reconstructed image, the output of the decoder. The supervised loss uses only the labeled portion of the training subset. To represent this term of the objective function we choose the cross-entropy loss. The supervised loss uses only the labeled portion of the training subset. To represent this term of the objective function we choose the cross-entropy loss. The minimization leads to the optimization of the parameters associated to the Encoder denoted by θ_E and the Dense Layer (θ_{D-L}) for the purpose of reducing the classification error, by relying on the supervised signals of the backward pass. The loss is expressed as follows:

$$l_{sup}\left(y_i, f\left(x_i + \zeta_i; \theta_E, \theta_{D-L}\right)\right) = -\sum_{j=0}^{C-1} y_{ij} log(f_j\left(x_i + \zeta_i; \theta_E, \theta_{D-L}\right)) \tag{2}$$

Where y_i is the ground truth label, $f\left(x_i + \zeta_i; \theta_E, \theta_{D-L}\right)$ refers to the prediction $\widetilde{y}_i$ of dense layer that uses the latent space coming from the Encoder, and ζ_i denotes the additive gaussian noise. As forementioned summing the supervised and unsupervised terms gives us the objective function:

$$l_{ssl} = l_{sup}\left(y_i, f\left(x_i + \zeta_i; \theta_E, \theta_{D-L}\right)\right) + \lambda . l_{unsup}\left(x_i, \widetilde{x}_i\right) \tag{3}$$

Here λ is a hyper-parameter used to control the degree to which the unsupervised loss contributes to the overall loss.

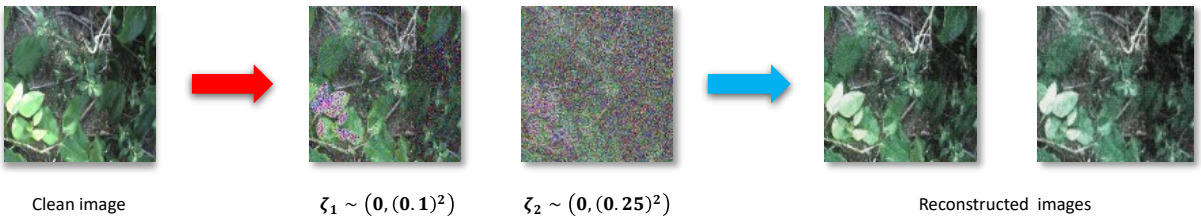


Figure 2: Reconstruction from noisy input images using our proposed auto-encoder (shown in Figure 1), the first gaussian noise ζ_1 was sampled from a normal distribution of a mean μ_1 equal to zero and a standard deviation σ_1 equal to 0.1, as for the second gaussian noise ζ_2 the parameters of its normal distribution are the following: μ_2 equal to zero and σ_2 equal to 0.25

3.3. ConvNeXt Encoder

ConvNeXt is a family of pure convolutional neural networks, developed through the progressive process of modernizing standard resnets (He et al., 2015), toward the design of transformers by exploring the design spaces and limits that can be reached by relying only on convnet modules (Liu et al., 2022). The series of steps adopted during the modernization process can be summarized in the following way:

First a change of the compute stage ratio, meaning changing the computation distribution across stages, by specifying the number of blocks that will be attributed to each stage. In a multi-stage design the feature map resolution changes depending on the stage.

The second step consists in replacing the standard resnet stem cell, with a 4×4 non-overlapping convolution layer called patchify stem. This design choice was motivated by the patchify layer of hierarchical vision transformers.

In the third step, a special case of grouped convolution was adopted, called depth-wise convolution, (Xie et al., 2016; Howard et al., 2017; Chollet, 2017) this is akin to the weighted sum operation in self-attention, as it operates on a per-channel basis, by mixing information only along the spatial dimension. The combination of depth-wise convolution, and 1×1 convolution allows for a separation between the operations of mixing information across the spatial dimension and mixing information across the channel dimension, which is an important principle shared by vision transformers.

The fourth step was inspired by the fact that in a transformer block (Dosovitskiy et al., 2020; Vaswani et al., 2017) an inverted bottleneck is created, due to the hidden dimension of the MLP, which is four times wider than the input dimension. This design is then used as part of the modernization process.

The fifth step is related to the implementation of the inverted bottleneck, which also follows the design of hierarchical vision transformers, where we can see that the Multi-head Self-Attention (MSA) module is positioned before the MLP layers (Dosovitskiy et al., 2020; Vaswani et al., 2017), therefore the depth-wise convolution with a large size kernel 7×7 will be positioned before the dense 1×1 convolution layer.

In the sixth step the ReLU function (Nair and Hinton, 2010) was replaced with the Gaussian Error Linear Unit (GELU) (Hendrycks and Gimpel, 2016), as it is the activation function used in the most advanced transformers. Also the number of activations per block was reduced to one, following this step the Batch Normalization (BN) layer (Ioffe and Szegedy, 2015) was substituted with Layer Normalization (LN) (Ba et al., 2016), which is the alternative choice of normalization in transformers, and the number of normalization layers used per block was also reduced to a single layer.

For the last step separate downsampling layers were used instead of relying on the residual block at the start of each stage. Before each downsampling layer an LN layer is added. Regarding the patchify stem the LN layer is added afterwards. This principle design is also used in hierarchical vision transformers.

The modernized family of ConvNeXt rivals vision transformers in terms of scalability and accuracy. For our proposed auto-encoder we have used the variation model called ConvNeXt-Base, to represent the encoder part of our joint neural networks, this variation has a compute stage ratio of (3, 3, 27, 3). Figure 3 shows the different elements (convolution layers, normalization layer, activation function) used in the ConvNeXt residual block, and what mainly changes depending on the stage is the number of kernels used on the level of convolution layers, and as we go along the

depth of the encoder from one stage to another, the number of channels increases, which results in an increase of computation inside the residual block (the encoder is depicted in Figure 4).

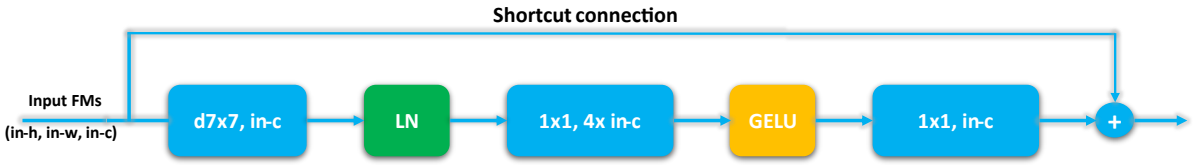


Figure 3: A ConvNeXt residual block (CB), with an input representing feature maps of resolution (in-h, in-w) and a number of channels equal to in-c. The number of kernels used in the depth-wise convolution layer and the last 1×1 convolution layer is equal to in-c, and as for the middle 1×1 convolution layer, the number of kernels is equal to four times in-c

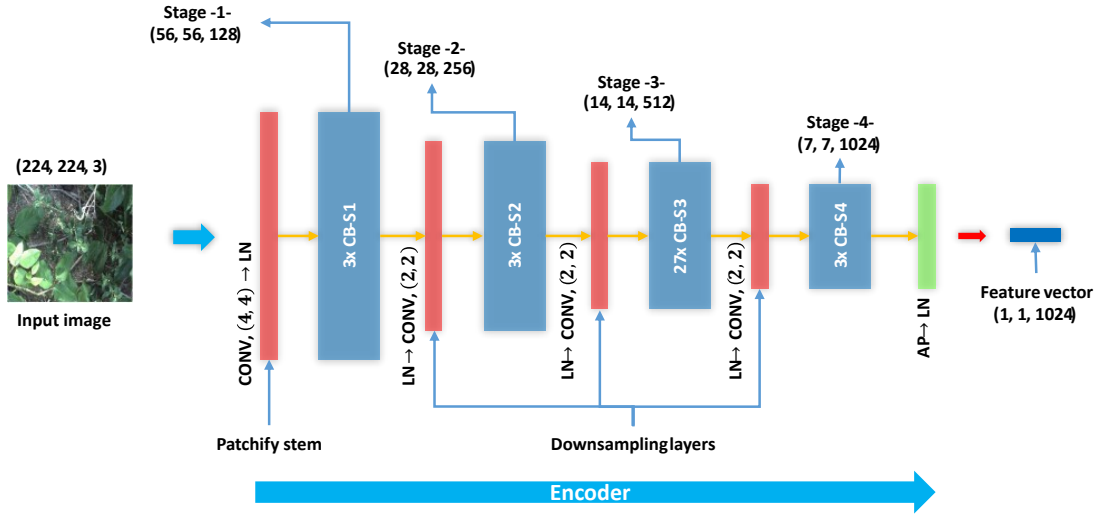


Figure 4: ConvNeXt-Base Encoder, which has a compute stage ratio of (3, 3, 27, 3) corresponding to the number of ConvNeXt residual blocks (CBs) belonging to each stage, and their outputs are used through the skip-connections to connect the encoder to the decoder (see Figure 5 for more details). CB is depicted in Figure 3

3.4. Decoder

As for the network representing the decoder part of our proposed auto-encoder, we have also adopted a multi-stage design, as it is illustrated in Figure 1, which will then allow us to utilize the feature maps from all stages of the ConvNeXt-Base encoder during the reconstruction process. The feature maps on the encoder side are retrieved through residual and non-residual skip-connections. The residual skip-connections provide the decoder with direct access to the outputs of the residual blocks at any given stage of the encoder. The non-residual skip-connections provide the decoder with direct access to the outputs of the downsampling layers, including the patchify stem. The number of skip-connections used in our proposed architecture is equal to 40. 36 residual skip-connections corresponding to the ConvNeXt residual blocks, and 4 non-residual skip-connections corresponding to the downsampling layers and the patchify stem. We chose these two types of skip-connections to capture the detailed information from the encoder path, and we introduced the intermediate hidden representations (the encoder feature maps) in a way that it takes into

consideration the symmetry between the encoder and the decoder, and also the fact that along the decoder path, we will go from low dimensional representations to high dimensional representations during the reconstruction process. Figure 5 shows the skip-connections connecting one stage of the decoder to its corresponding stage of the encoder. Similar to the design of the ConvNeXt-Base encoder, we also used separate upsampling layers at the beginning of each decoder stage. Adding these layers is very important for the structure of the decoder, as well as for the process of incorporating the information coming through the skip-connections at each stage of the decoder.

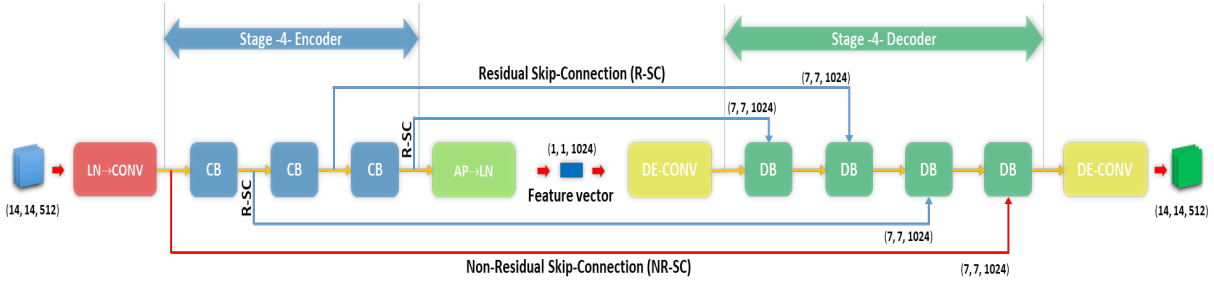


Figure 5: View of the skip-connections between stage 4 of the encoder and stage 4 of the decoder, where DB and CB denote the de-convolution block and the ConvNeXt residual block. The downsampling layer is preceded with layer normalization (LN → CONV), as for the adaptive average pooling, normalization is applied afterwards (AP → LN). The two separate de-convolution layers (DE-CONVs) on the decoder side represent the upsampling layers

In addition to the upsampling layers, every decoder stage has a specific number of de-convolution blocks associated to it. The number of these blocks is equal to the number of skip-connections connecting the stage to the encoder path. Each de-convolution block takes as inputs the encoder representations received from its corresponding skip-connection and the output of the previous de-convolution block or the separate upsampling layer. The source of this second input depends on where the de-convolution block is positioned in the stage. First and second inputs are then combined using element-wise summation, which results in a mixing between the spatial information of the decoder and the encoder. Following this step a 1×1 de-convolution layer of stride 1 is applied, with a number of kernels equal to the number of channels of the summed inputs. Due to the element-wise summation the dimensions of representations don't change inside the de-convolution block. Afterwards the output of the 1×1 de-convolution layer is normalized through layer normalization, and in the last step an activation function is applied to obtain the overall output of the block.

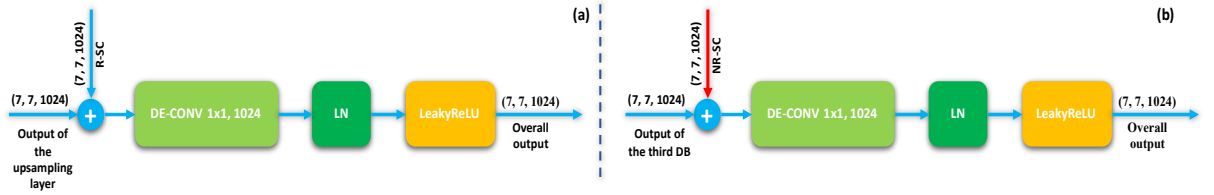


Figure 6: Detailed view of the first de-convolution block (a) and the fourth de-convolution block (b) belonging to stage -4- of the decoder

Regarding the activation functions, applied element-wise on the decoder side, we used for the stages 4, 3, and 2 the activation function LeakyReLU (Maas et al., 2013), which was introduced to alleviate potential problems related to the hard 0 activation of the ReLU function, by allowing a

small, non-zero gradient when the unit is saturated. As for stage 1, we used the activation function ELU (Clevert et al., 2015) which can help in solving the vanishing gradient problem and could speed up the learning because of its characteristics.

At the end of the decoder, we added a final 4×4 de-convolution layer of stride 4, and a number of kernels equal to 3, followed by a sigmoid function applied element-wise, in order to obtain the output image. We have used the sigmoid function because the pixel values of each input image will be scaled between 0 and 1, through the division of all three channels by 255, the full decoder architecture is summarized in Figure 7.

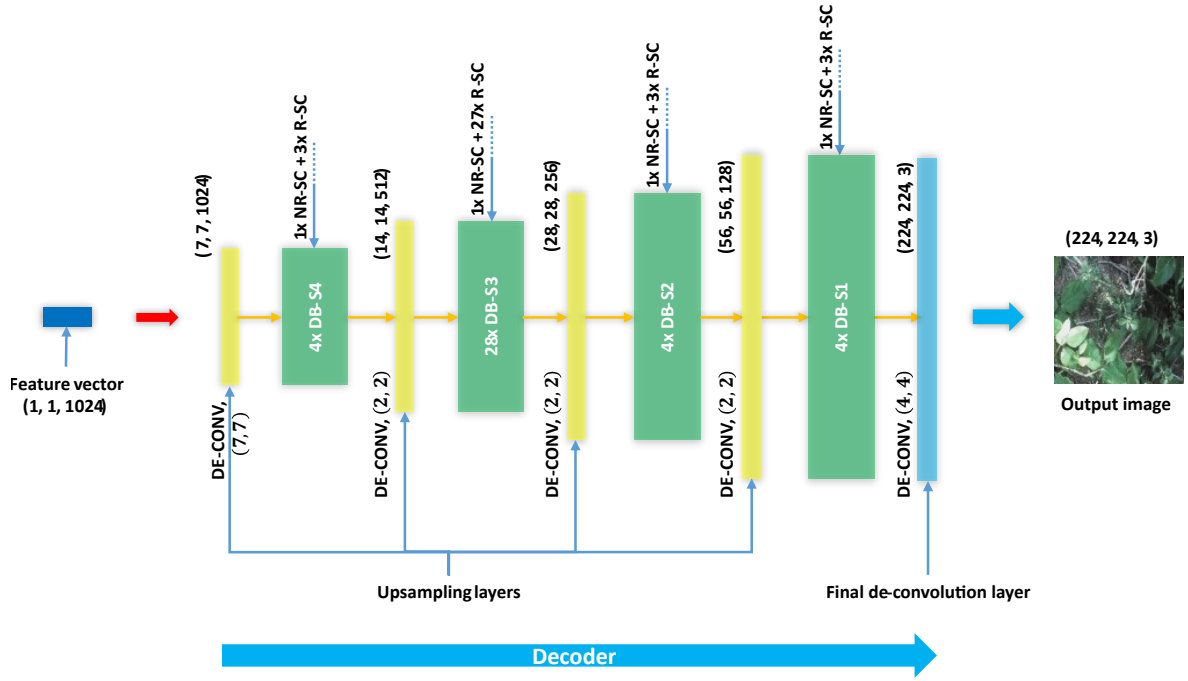


Figure 7: Decoder network, with a multi-stage design to ensure the overall symmetric structure of the proposed auto-encoder and the integration of skip-connections. An overview of the encoder-decoder model is shown in Figure 1. Reconstructed outputs from noisy inputs are presented in Figure 2

For the semi-supervised setting an additional Dense-Layer (D-L, see Figure 1) is added, with a number of units equal to the number of classes we want to identify. The layer takes as input the feature vector (output of the ConvNeXt Encoder) and it is followed with a SoftMax function applied element-wise. During the reconstruction process, we added gaussian noise to the input images. This choice of reconstructing from noisy (corrupted) inputs was motivated by the smoothness assumption of semi-supervised learning, which implies that a realistic perturbation of the input should not change the output label of the model (Yang et al., 2021).

4. Experiments and results

This section first introduces the experimental setup and the key aspects of the evaluation procedure, then comprehensively assess the performance of the proposed semi-supervised approach for plant classification, particularly focusing on weed datasets. We compare our method with

state-of-the-art supervised models, that employ a similar range of number of parameters to show the benefit of the semi-supervised approach. Additionally, we conduct experiments to compare the proposed method with similar semi-supervised techniques on CIFAR10 dataset.

4.1. Training and evaluation procedure

4.1.1. Selected deep learning models for comparison

In addition to the ConvNeXt-Base supervised model, we have also chose two types of models belonging to two different families of neural networks, based on their performance and the availability of weights pre-trained on the large dataset ImageNet-22K, which is a superset of ImageNet-1K, and it is comprised of 14,197,122 labeled images divided into 21841 $\approx$ 22K classes. Using these pre-trained weights as a starting point for training on the target datasets (DeepWeeds and 4-Weeds), is beneficial in terms of performance and reducing the required time to train the models, compared to the method of initializing the weights randomly. These two models are Efficientnet-V2-L (Tan and Le, 2021)and ViT-B-16 (Dosovitskiy et al., 2020), the first type is part of the improved models of the family Efficientnet-V2, that are optimized to have a faster training speed and efficient parameters, and the second type part of the vision transformer family.

Model	**# Parameters**	**Top-1 @cc (%)**	**Architecture**
ConvNeXt-Base	88.59M	84.06	Convolutional
Efficientnet-V2-L	118.58M	85.80	Convolutional
ViT-B-16	86.46M	81.07	Vision Transformer

Table 1: Models performance on ImageNet-1k

4.1.2. Parameters setup

Regarding the implementation of the deep learning algorithms and the running of the experiments, we used the machine learning framework PyTorch **1.10.2**, along with the library PyTorch-Ignite **0.4.8**, and the library timm **0.5.4**, and in terms of hardware specifications, we used the regional computing cluster CaSciModOT, and through this cluster we were able to access three GPUs of the type Nvidia Tesla V100 32GB and the AMD 7302 CPU. In order to ensure an objective comparison and evaluation of the performances, that are achieved by the deep learning models, the same optimal hyperparametric parameters were used consistently in all the conducted experiments. The initial learning rate was set at 0.01, decreasing by a factor of 0.9 every 10 epochs. The stochastic gradient descent (SGD) algorithm was used to update model parameters, with momentum and damping set to zero. With regard to the use of pre-trained weights, the dense ImageNet-22k layer was replaced by a layer containing a number of units corresponding to the classes specific to the dataset used during training (9 units for DeepWeeds, 4 units for 4-Weeds). With regard to the characteristics of the gaussian noise added during the semi-supervised training of the ConvNeXt-Base models, the standard deviation $\boldsymbol{\sigma}$ was set to 0.1, while the mean $\boldsymbol{\mu}$ was set to zero, and to initialize the decoders, pre-trained weights on ImageNet-1K were used. As for the hyperparameter $\boldsymbol{\lambda}$ that balances the unsupervised and supervised terms of the objective function, it was set to 1. For generating unlabeled data, a single transformation was applied to the **60%** labeled portion of the training subset, involving rotations ranging from -180°to +180°. The resulting images are then de-labeled and used during the semi-supervised learning. We should also note that when reducing

the size of the labeled portion of the training subset from **60%** to **20%**, the labels of the unsampled images are removed, and added to the unlabeled portion of the training subset.

4.1.3. Evaluation procedure

Two metrics were selected for performance evaluation of the deep learning models, namely: F1-Score and Accuracy (Sokolova et al., 2006; Tharwat, 2018; Opitz and Burst, 2019). The first metric represents the harmonic mean of two metrics: 1. Precision that denotes the proportion of positive predictions that are correctly classified relative to the total number of positive predictions ($FP + TP$). 2. Recall which is the proportion of positive predictions that are correctly classified relative to the real number of positives ($FN + TP$). As for the second metric, it is defined as the ratio between correctly classified predictions and the total number of predictions ($FN + FP + TP + TN$). The equations that define the metrics are expressed as follows:

$$\boldsymbol{F1\text{-}Score} = 2 \times ((Precision \times Recall) \,/\, (Precision + Recall)) \tag{4}$$

$$\boldsymbol{Precision} = (TP) \,/\, (TP + FP)\,, \boldsymbol{Recall} = (TP)\,(TP + FN) \tag{5}$$

$$\boldsymbol{Accuracy} = (TP + TN) \,/\, (TP + TN + FP + FN) \tag{6}$$

Throughout the experiments that we conducted, we used the 5-fold stratified cross-validation (Varma and Simon, 2006; Arlot and Celisse, 2010; Raschka, 2018). Its goal is to provide an accurate estimate of how well a given model will perform in practice on new unseen data samples. Stratified sampling was used to ensure that each fold will be approximately representative of the original dataset in terms of class distribution. We initially partitioned the datasets (DeepWeeds and 4-Weeds) in the following manner: **60%** for training, **20%** for validation and the remaining **20%** for testing, with five rounds of cross-validation. This means that for each model type, a total number of five models are trained on different partitions of the data, and we have used the validation subset for the selection of the optimal hyperparameters. During the learning process, no data augmentation techniques were employed to increase the size of the labeled portion. On the contrary, as mentioned previously, we avoided such augmentation in order to accurately assess its impact on model performance. We have carried-out three sets of experiments with variable amount of labeled data. For the first set we used **60%** of the labeled training samples. For the second set we used **40%** of the labeled samples. For the last set of experiments, we used only **20%** of the labeled samples. All models were trained for a duration of time equal to 60 epochs, and for each model type belonging to a given family, the best 5 models in terms of their performance on the validation subset were saved, because of the performed 5 round cross-validation, then those models are evaluated on the test subset, to obtain the final average performance reported here.

4.2. Experiments on DeepWeeds dataset

4.2.1. DeepWeeds description

For the training and evaluation of the learning algorithms we have used the DeepWeeds dataset first (Olsen et al., 2019), which consists of in-situ images of eight different weed species (target classes) native to eight locations in northern Australia, and various plant life (negative class) indigenous to those locations. The images were collected from June 2017 to March 2018, under realistic environmental conditions, to capture the challenges related to the classification of weed species, stemming from several factors of variation such as: geographical and seasonal variation of plants, dynamic complex backgrounds, illumination variation that results in highly dynamic

range scenes with bright reflectance and dark shadows. With the intent of attaining the required variability and generality of the dataset, over 1000 images were gathered of each target species, with approximately a 50:50 split of positive to negative class images from each location, leading to a total of 8403 positive images of weed species and 9106 negative images of neighboring flora and non-target backgrounds, all the images of the DeepWeeds dataset were labeled by botanists of Figure 8 and 9.

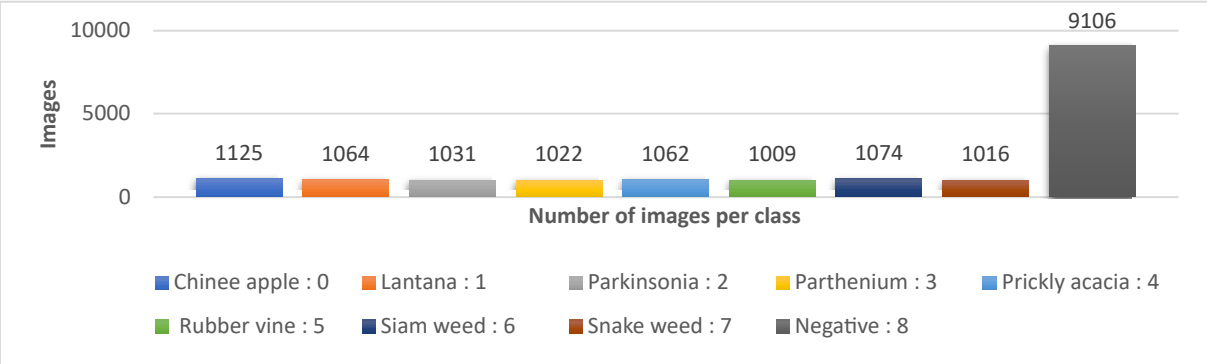


Figure 8: Class distribution of the DeepWeeds dataset

Figure 9: Sample images of the weed species and the non-target flora

4.2.2. Classification results on DeepWeeds

The results of evaluation of the different types of the deep learning models are summarized in Figure 10 and 11, based on the two metrics Accuracy and F1-Score. From these results we can observe that the best average performance is achieved by ConvNeXt-Base-SSL, which refers to the models trained in semi-supervised way through our proposed auto-encoder and the objective function composed of the supervised and unsupervised terms. Moreover, in the case where the size of the labeled training subset is the smallest, i.e., **20%**, we were able to achieve a strong classification performance, due to the utilization of unlabeled data, and this performance is greater than 90% for both metrics, and has an important margin compared to the best supervised performance obtained by ViT-B-16, with a difference of 1.23% in Accuracy and 1.9% in F1-Score. Furthermore, the difference in performance is even more significant, when the comparison is carried out with respect to Efficientnet-V2-L, where we can observe that we have a difference of 6.88% and 9.68% in Accuracy and F1-Score respectively. Regarding the variation of labeled data used, the supervised

models of Efficientnet-V2-L demonstrated a significant drop in performance when the number of labeled images used during training was limited. As we reduced the labeled training portion from **60%** to **20%**, these models experienced a decline of 6.14% in Accuracy and 8.42% in F1-Score. In contrast the semi-supervised models showed a small decrease in performance, with a drop of 2.84% in Accuracy and 3.48% drop in F1-Score, which proves that the models trained through our proposed auto-encoder and objective function are more resilient to the change in the size of the labeled training subset.

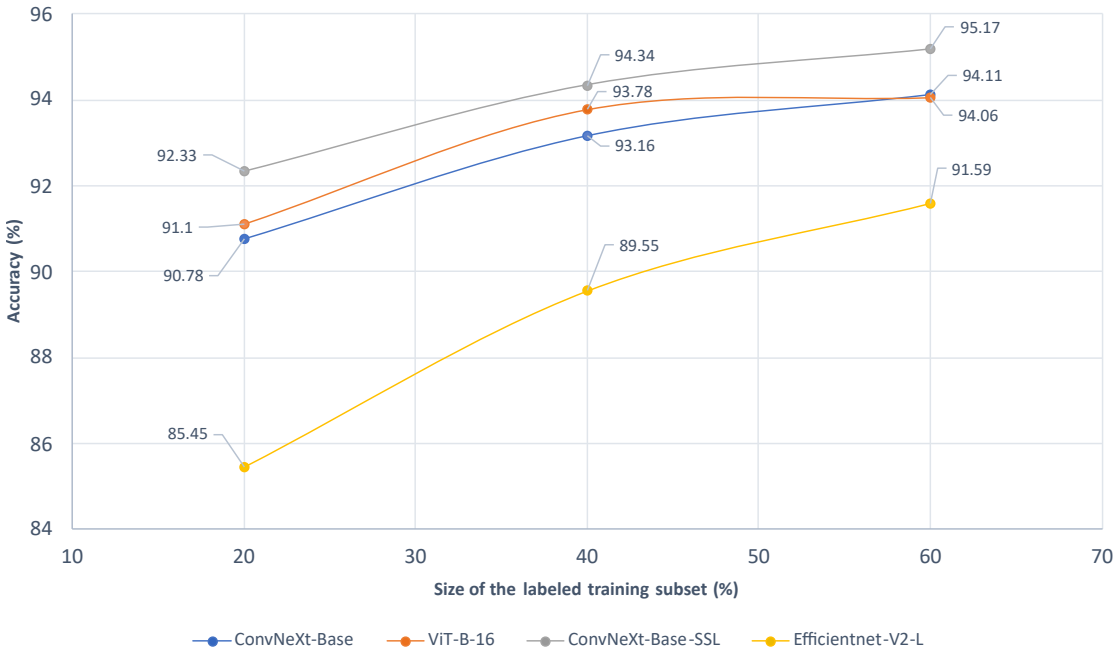


Figure 10: **Average** performance of the deep learning models on the **test subset**, based on the **Accuracy** metric, with different sizes of the labeled training subset

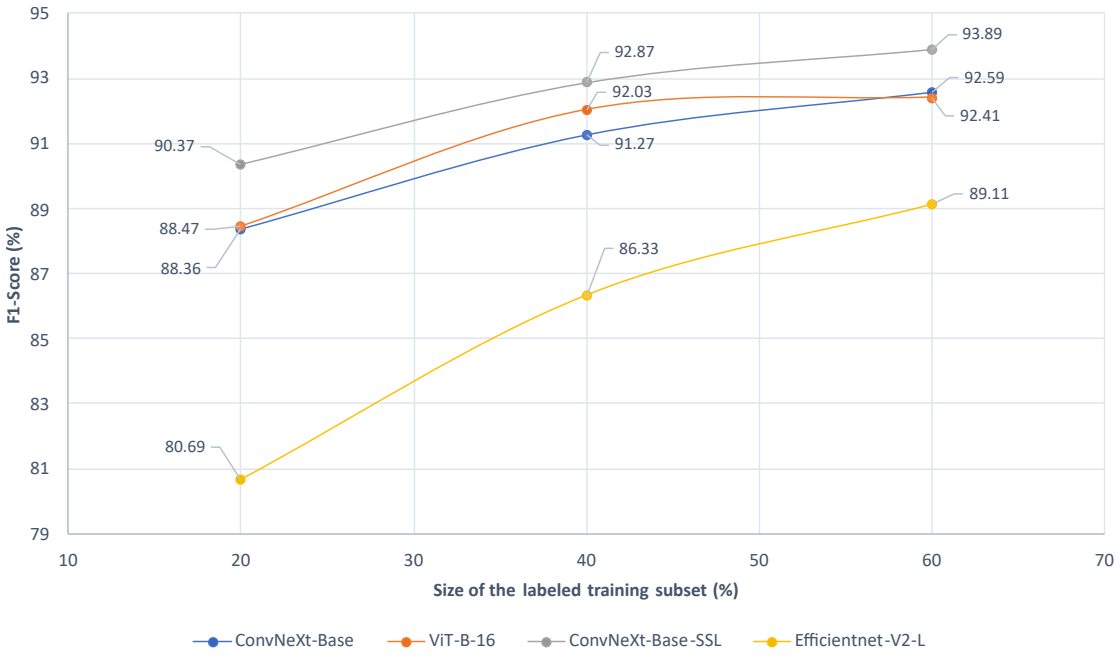


Figure 11: **Average** performance of the deep learning models on the **test subset**, based on the **F1-Score** metric, with different sizes of the labeled training subset

4.2.3. Noise influence

In this step, a series of evaluations were carried out, using the best performing deep learning models, saved from the completed three sets of experiments. The evaluations were performed by using noisy images of the test subset, resulting from the added gaussian noise, sampled from a normal distribution with a fixed mean $\boldsymbol{\mu}$, and a varying standard deviation $\boldsymbol{\sigma}$. Figure 12 shows the classification results of the models trained using only **20%** of labeled images. From these results we can observe that introducing additive gaussian noise to the test subset has a negative effect on the supervised models, represented by a significant decrease in performance, which is correlated with the increase in the value of standard deviation $\boldsymbol{\sigma}$. For instance, the performance of ViT-B-16 drops by 8.43% in Accuracy and by 11.87% in F1-Score, when $\boldsymbol{\sigma}$ is increased to 0.125. The performance drop of Efficientnet-V2-L was much more significant comapred to ViT-B-16 and ConvNeXt-Base. The same negative correlation between classification performance and the increase in noise importance is observed, regarding the supervised models trained on a larger number of labeled images (see Figure 13 and 14). On the other hand, the semi-supervised models provide a stable classification performance, relative to the increase of $\boldsymbol{\sigma}$. In the case where we have a smaller or higher number of labeled training images, which proves that the models trained through our proposed method, are robust to the presence of noise, which leads to better generalization on new unseen samples.

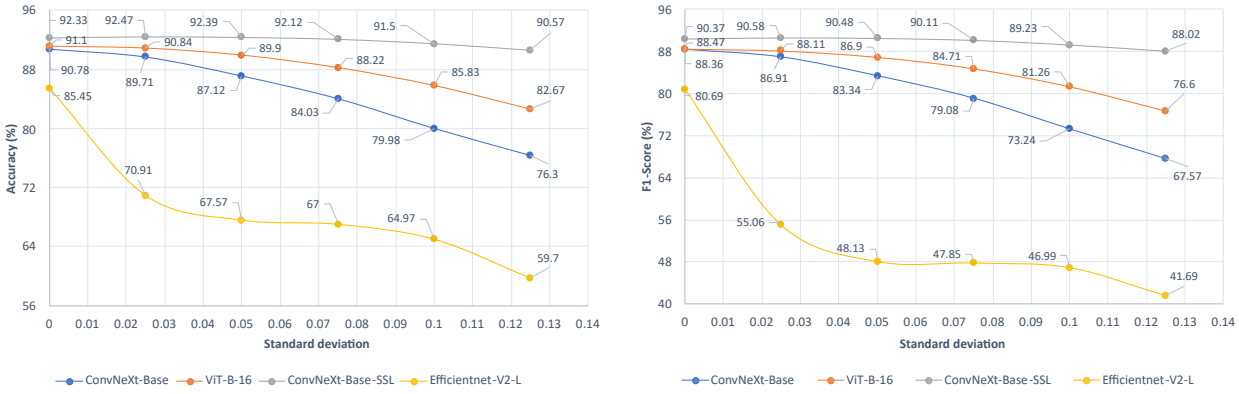


Figure 12: **Average performance** of deep learning models on the **test subset**, regarding the third data partitioning : **20%** labeled training subset, **20%** validation, **20%** test, with added gaussian noise to the inputs during inference, the x axis represents the standard deviation $\boldsymbol{\sigma}$ values associated with the gaussian noise (the mean $\boldsymbol{\mu}$ is fixed and its equal to zero)

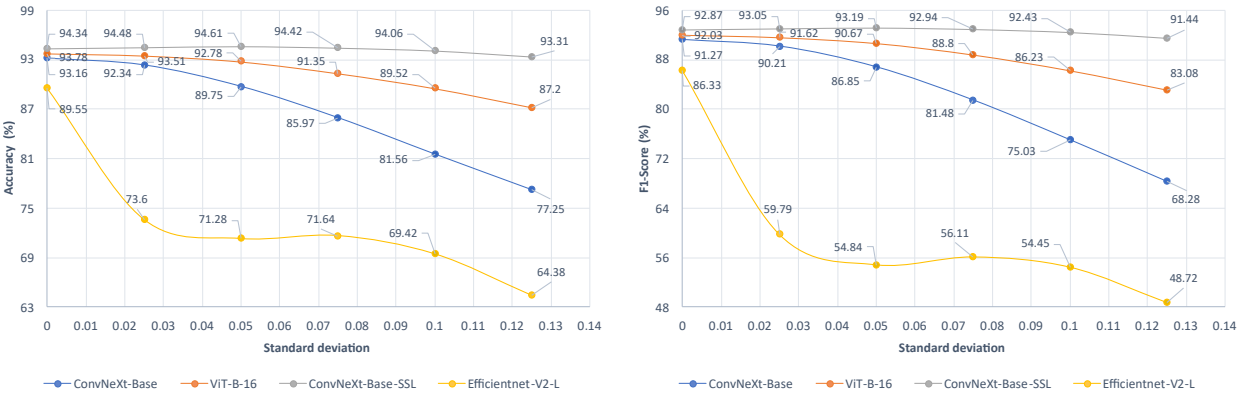


Figure 13: **Average performance** of deep learning models on the **test subset**, regarding the second data partitioning : **40%** labeled training subset, **20%** validation, **20%** test, with added gaussian noise to the inputs during inference

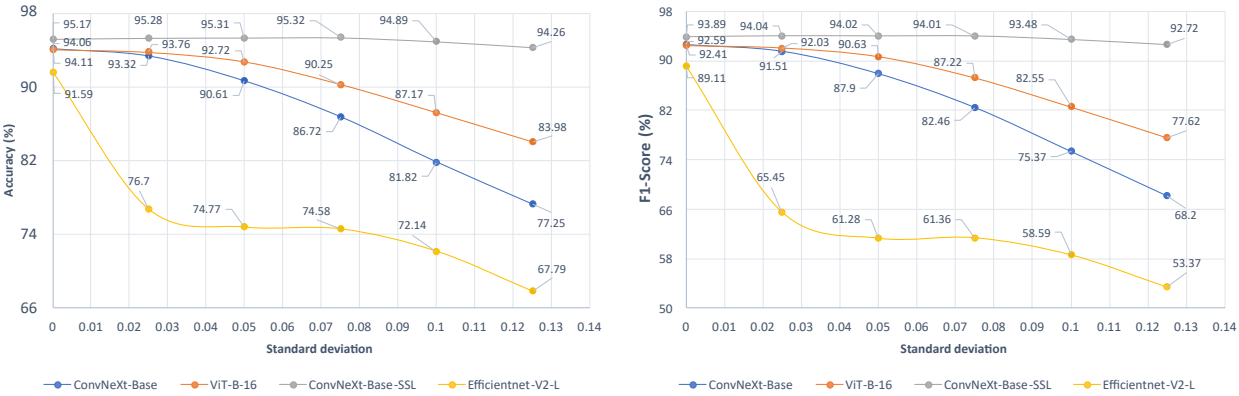


Figure 14: **Average performance** of deep learning models on the **test subset**, regarding the second data partitioning : **60%** labeled training subset, **20%** validation, **20%** test, with added gaussian noise to the inputs during inference

4.2.4. *Skip-Connections relevance*

Following the series of evaluations with respect to noise influence on the deep learning models, we performed further experiments with the aim of highlighting the importance of skip-connections to the overall architecture of our semi-supervised approach. For these experiments, we removed all the skip-connections connecting the ConvNeXt Encoder to the Decoder model. We kept the same training conditions as described previously and we have used **20%** of the labeled samples during the learning process. The differences in the average performance of ConvNeXt-Base-SSL, resulting from the removal of skip-connections are shown in Table 2.

Model	**Labeled data (%)**	**Accuracy (%)**	**F1-Score (%)**	**Reconstruction Error, MSE (10^{-3})**
ConvNeXt-Base-SSL **with SCs (# 40)**	**20**	92.33	90.37	5.30
ConvNeXt-Base-SSL **without SCs**	**20**	91.09	88.66	40

Table 2: **Average performance** of **ConvNeXt-Base-SSL** on the **test subset**,**with** and **without Skip-Connections** using only **20%** of labeled data during training

The impact on the reconstruction error is considerably significant, which is represented by a factor of increase of **7.53** in comparison to the result obtained when all skip-connections are utilized. The performance differences on the two metrics, Accuracy and F1-Score, are both important where we lose **1.24%** and **1.71%**, respectively, due to the absence of skip-connections.

4.3. *Experiments on 4-Weeds dataset*

4.3.1. *4-Weeds description*

For the purpose of carrying out an extended assessment of the learning algorithms, with a particular focus on the two model types: ConvNeXt-Base, and ConvNeXt-Base-SSL, we proceeded with the training and evaluation on the 4-Weeds dataset (Aggarwal et al., 2022). We used the same settings applied during the previous three sets of experiments completed on DeepWeeds with all the skip-connections utilized, in terms of optimal hyperparameters, stratified cross-validation, data partitioning, and unlabeled data generation. And with respect to the 4-Weeds dataset, we used the latest updated version, which consists of 618 labeled images of four commonly weed species found in corn and soybean, these images were acquired under complex field conditions during summer months at Purdue's University ACRE farm, as for winter months at Purdue's University greenhouse.

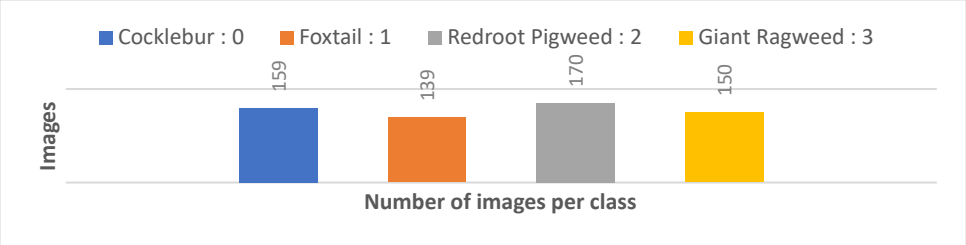


Figure 15: Class distribution of the 4-Weeds dataset

Figure 16: Images of the four species belonging to 4-Weeds

4.3.2. Classification results on 4-Weeds

The results of evaluation using 4-Weeds dataset are depicted in Figure 17, through these results we can observe that ConvNeXt-Base-SSL achieved a strong classification performance (F1-Score of **95.02%**), with a very limited size of labeled training images, which is equal to **20%**. This performance is greater than the classification results attained by ConvNeXt-Base in terms of both metrics, with a significant margin of 8.58% in Accuracy and 9.51% in F1-Score. As for when the size of the labeled training subset is higher, we can observe that we continue to have an important margin between the semi-supervised and supervised performances, which is approximately equal to 2% for both metrics. These results demonstrate that the semi-supervised models are more capable of producing a consistent and high classification performance, despite of the limitation in terms of labeled training samples, which is in accordance with what we have seen in the experiments conducted using DeepWeeds.

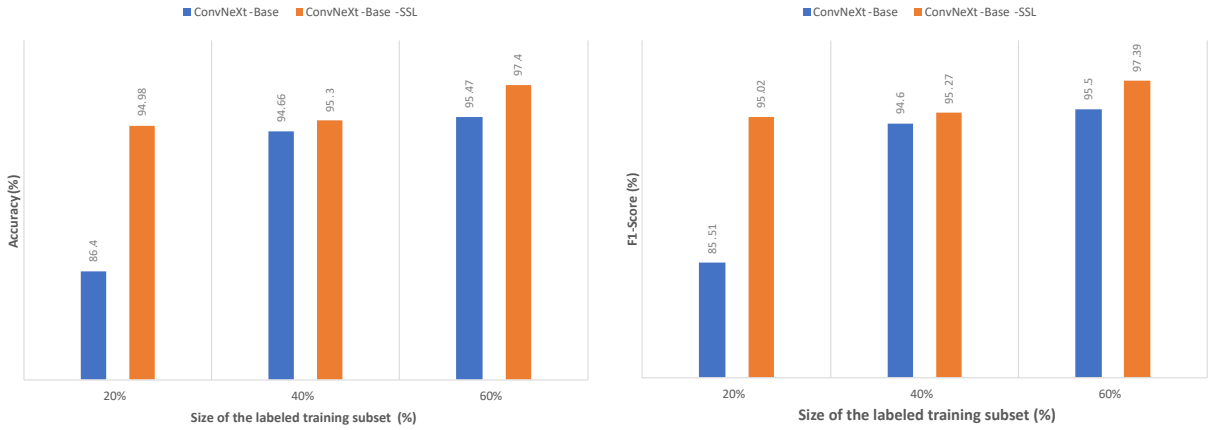


Figure 17: **Average performance** of deep learning models on the **test subset**, corresponding to different sizes of the labeled training subset

4.4. Complementary experiments on CIFAR-10 dataset

In this subsection, we build upon our previous findings by providing additional experimental results on the CIFAR-10 dataset. CIFAR-10 is a collection of 60000 32x32 RGB images, divided into 10 categories (classes) with 6000 images per category. The dataset is split into a training subset of 50000 images and a test subset of 10000 images. We used the same hyperparameters and initialization method for the encoder-decoder parameters (weights) as in the training on DeepWeeds and 4-Weeds. However, in this case, we did not apply the cross-validation technique. We have used random seeds for the stratified sampling of 1000, 2000, and 4000 labeled samples, respectively.

The results of evaluation on the test subset of CIFAR-10 are shown in Table 3, along with a comparison to the consistency regularization-based methods mentioned earlier, in addition to two semi-supervised methods namely: FixMatch (Sohn et al., 2020), and SelfMatch (Kim et al., 2021). The first combines pseudo labeling of high-confidence predictions with consistency regularization based on weak and strong augmentations, the second utilizes contrastive learning for pre-training, and adopts the learning strategy of FixMatch for the semi-supervised training.

Semi-supervised model	**1k labels**	**2k labels**	**4k labels**
π-model (Laine and Aila, 2016)	68.35±1.20	82.43±0.44	87.64±0.31
Temporal ensembling (Laine and Aila, 2016)	76.69±1.01	84.36±0.39	87.84±0.24
Mean teacher (Tarvainen and Valpola, 2017)	81.22±0.31	82.43±0.20	87.64±0.27
Dual student (Ke et al., 2019)	84.26±0.45	88.53±0.14	90.35±0.12
FixMach (Sohn et al., 2020)	-	-	95.74±0.05
SelfMatch (Kim et al., 2021)	-	-	95.94±0.08
ConvNeXt-Base-SSL	94.77±0.72	97.30±0.30	97.97±0.09

Table 3: Performance of semi-supervised models over 5 runs on the test subset of CIFAR-10. The results of comparison for the consistency regularization-based methods were reported from (Ke et al., 2019). The results of FixMatch and SelfMatch were reported from (Kim et al., 2021)

The results on CIFAR-10 show that ConvNeXt-Base-SSL achieves the best classification performance, ranging from 94.77% to 97.97%. This is true for all cases, including when we have 1000 and 2000 labeled samples, as well as when we have 4000 labeled samples. ConvNeXt-Base-SSL outperforms Dual Student by 10.51% when we only have 1000 labeled samples, and achieves higher performance than FixMatch and SelfMatch, with an approximate margin of 2%, when using 4000 labeled samples. This further validates the direction and steps we took in developing our approach.

5. Conclusions

In this paper, we have proposed a method based on the semi-supervised learning paradigm, for the purpose of developing highly performing and reliable deep learning models, in terms of robustness and generalization, to accurately identify and recognize weeds species, under real-world conditions, from ground-based images. To this end, we developed a new deep learning architecture, which consists of the modernized encoder ConvNeXt-Base, and the thoroughly designed decoder,

to enable the successful targeted incorporation of consistency regularization and thereby the utilization of both types of data, labeled and unlabeled during the learning process. The extensive series of experimentations showed that the models trained through our proposed method provide a stable and strong classification performance, despite the very limited amounts of labeled samples available during training. In comparison, state-of-the-art supervised deep learning models, were affected negatively by the decrease in the number of the labeled training samples and the presence of noise during inference. Moreover, the experiments highlight the performance effectiveness of our proposed method compared to similar semi-supervised learning approaches. In future work, we intend to improve furthermore the generalization ability of the deep semi-supervised models and extend our method to other computer vision applications.

Credit Author Statement

Farouq Benchallal: Conceptualization, Methodology, Software, Writing – review & editing, Writing – original draft preparation, Writing – original draft, Visualization, Investigation, Validation.

Adel Hafiane: Conceptualization, Supervision, Writing – original draft preparation, Project administration, Writing – review & editing, Methodology.

Nicolas Ragot: Conceptualization, Supervision, Writing – original draft preparation, Project administration, Writing – review & editing, Methodology.

Raphaël Canals: Conceptualization, Supervision, Writing – original draft preparation, Project administration, Writing – review & editing, Methodology.

Acknowledgments

This work was carried out as a part of DESHERBROB project funded by Region Centre-Val de Loire. We gratefully acknowledge its support.

References

Aggarwal, V., Ahmad, A., Etienne, A., Saraswat, D., 2022. 4weed dataset: Annotated imagery weeds dataset. arXiv:2204.00080.

Ahmad, A., Saraswat, D., Aggarwal, V., Etienne, A., Hancock, B., 2021. Performance of deep learning models for classifying and detecting common weeds in corn and soybean production systems. Computers and Electronics in Agriculture 184. doi:10.1016/j.compag.2021.106081.

Aktar, W., Sengupta, D., Chowdhury, A., 2009. Impact of pesticides use in agriculture: Their benefits and hazards. Interdisciplinary Toxicology 2, 1–12. doi:10.2478/v10102-009-0001-7.

Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Statistics Surveys 4, 40–79. doi:10.1214/09-SS054.

Ba, J.L., Kiros, J.R., Hinton, G.E., 2016. Layer normalization URL: http://arxiv.org/abs/1607.06450.

Bah, M.D., Hafiane, A., Canals, R., 2018. Deep learning with unsupervised data labeling for weed detection in line crops in uav images. Remote Sensing 10. URL: https://www.mdpi.com/2072-4292/10/11/1690, doi:10.3390/rs10111690.

Chapelle, O., Schölkopf, B., Zien, A. (Eds.), 2006. Semi-Supervised Learning. The MIT Press. URL: https://doi.org/10.7551/mitpress/9780262033589.001.0001, doi:10.7551/mitpress/9780262033589.001.0001.

Chauhan, B.S., 2020. Grand challenges in weed management. doi:10.3389/fagro.2019.00003.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Clevert, D.A., Unterthiner, T., Hochreiter, S., 2015. Fast and accurate deep network learning by exponential linear units (elus) URL: http://arxiv.org/abs/1511.07289.

Cong, R., Lei, J., Fu, H., Cheng, M.M., Lin, W., Huang, Q., 2019. Review of visual saliency detection with comprehensive information. IEEE Transactions on Circuits and Systems for Video Technology 29, 2941–2959. doi:10.1109/TCSVT.2018.2870832.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An image is worth 16x16 words: Transformers for image recognition at scale URL: http://arxiv.org/abs/2010.11929.

Gharde, Y., Singh, P.K., Dubey, R.P., Gupta, P.K., 2018. Assessment of yield and economic losses in agriculture due to weeds in india. Crop Protection 107, 12–18. doi:10.1016/j.cropro.2018.01.007.

Hasan, A.S., Sohel, F., Diepeveen, D., Laga, H., Jones, M.G., 2021. A survey of deep learning techniques for weed detection from images. doi:10.1016/j.compag.2021.106067.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition URL: http://arxiv.org/abs/1512.03385.

Heap, I., 2023. Current status of the international herbicide-resistant weed database. https://www.weedscience.org. Accessed on 2023-06-08.

Hendrycks, D., Gimpel, K., 2016. Gaussian error linear units (gelus) URL: http://arxiv.org/abs/1606.08415.

Homan, D., du Preez, J.A., 2021. Automated feature-specific tree species identification from natural images using deep semi-supervised learning. Ecological Informatics 66, 101475. URL: https://www.sciencedirect.com/science/article/pii/S1574954121002661, doi:https://doi.org/10.1016/j.ecoinf.2021.101475.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications URL: http://arxiv.org/abs/1704.04861.

Hu, C., Thomasson, J.A., Bagavathiannan, M.V., 2021. A powerful image synthesis and semi-supervised learning pipeline for site-specific weed detection. Computers and Electronics in Agriculture 190, 106423. URL: https://www.sciencedirect.com/science/article/pii/S0168169921004403, doi:https://doi.org/10.1016/j.compag.2021.106423.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift URL: http://arxiv.org/abs/1502.03167.

Kamilaris, A., Prenafeta-Boldú, F.X., 2018. Deep learning in agriculture: A survey. doi:10.1016/j.compag.2018.02.016.

Ke, Z., Wang, D., Yan, Q., Ren, J., Lau, R.W.H., 2019. Dual student: Breaking the limits of the teacher in semi-supervised learning URL: http://arxiv.org/abs/1909.01804.

Kerdegari, H., Razaak, M., Argyriou, V., Remagnino, P., 2019. Semi-supervised GAN for classification of multispectral imagery acquired by uavs. CoRR abs/1905.10920. URL: http://arxiv.org/abs/1905.10920, arXiv:1905.10920.

Khan, S., Tufail, M., Khan, M.T., Khan, Z.A., Iqbal, J., Alam, M., 2021. A novel semi-supervised framework for uav based crop/weed classification. PLoS ONE 16. doi:10.1371/journal.pone.0251008.

Kim, B., Choo, J., Kwon, Y.D., Joe, S., Min, S., Gwon, Y., 2021. Selfmatch: Combining contrastive self-supervision and consistency for semi-supervised learning URL: http://arxiv.org/abs/2101.06480.

Krizhevsky, A., 2009. Learning multiple layers of features from tiny images. URL: https://www.cs.toronto.edu/~kriz/cifar.html.

Laine, S., Aila, T., 2016. Temporal ensembling for semi-supervised learning URL: http://arxiv.org/abs/1610.02242.

Lecun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. doi:10.1038/nature14539.

Li, W., Hosseini Jafari, O., Rother, C., 2019. Deep object co-segmentation, in: Jawahar, C.V., Li, H., Mori, G., Schindler, K. (Eds.), Computer Vision – ACCV 2018, Springer International Publishing, Cham. pp. 638–653.

Liu, T., Jin, X., Zhang, L., Wang, J., Chen, Y., Hu, C., Yu, J., 2023. Semi-supervised learning and attention mechanism for weed detection in wheat. Crop Protection 174, 106389. URL: https://www.sciencedirect.com/science/article/pii/S0261219423002120, doi:https://doi.org/10.1016/j.cropro.2023.106389.

Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s URL: http://arxiv.org/abs/2201.03545.

Llewellyn, R., Ronning, D., Clarke, M., Mayfield, A., Consulting, A.M., Walker, S., 2016. IMPACT OF WEEDS ON AUSTRALIAN GRAIN PRODUCTION The cost of weeds to Australian grain growers and the adoption

of weed management and tillage practices Report for Grains Research and Development Corporation. URL: www.grdc.com.au/bookshop.
Maas, A.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier nonlinearities improve neural network acoustic models.
Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted boltzmann machines, in: Fürnkranz, J., Joachims, T. (Eds.), Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel, Omnipress. pp. 807–814. URL: https://icml.cc/Conferences/2010/papers/432.pdf.
Oerke, E.C., 2006. Crop losses to pests. doi:10.1017/S0021859605005708.
Oliver, A., Odena, A., Raffel, C., Cubuk, E.D., Goodfellow, I.J., 2018. Realistic evaluation of deep semi-supervised learning algorithms, in: Neural Information Processing Systems.
Olsen, A., Konovalov, D.A., Philippa, B., Ridd, P., Wood, J.C., Johns, J., Banks, W., Girgenti, B., Kenny, O., Whinney, J., Calvert, B., Azghadi, M.R., White, R.D., 2019. Deepweeds: A multiclass weed species image dataset for deep learning. Scientific Reports 9. doi:10.1038/s41598-018-38343-3.
Opitz, J., Burst, S., 2019. Macro f1 and macro f1 URL: http://arxiv.org/abs/1911.03347.
Pimentel, D., Zuniga, R., Morrison, D., 2005. Update on the environmental and economic costs associated with alien-invasive species in the united states. Ecological Economics 52, 273–288. doi:10.1016/j.ecolecon.2004.10.002.
Radicetti, E., Mancinelli, R., 2021. Sustainable weed control in the agro-ecosystems. doi:10.3390/su13158639.
Radoglou-Grammatikis, P., Sarigiannidis, P., Lagkas, T., Moscholios, I., 2020. A compilation of uav applications for precision agriculture. Computer Networks 172. doi:10.1016/j.comnet.2020.107148.
Raschka, S., 2018. Model evaluation, model selection, and algorithm selection in machine learning URL: http://arxiv.org/abs/1811.12808.
Rasmus, A., Valpola, H., Honkala, M., Berglund, M., Raiko, T., 2015. Semi-supervised learning with ladder networks URL: http://arxiv.org/abs/1507.02672.
Reedha, R., Dericquebourg, E., Canals, R., Hafiane, A., 2022. Transformer neural network for weed and crop classification of high resolution uav images. Remote Sensing 14. doi:10.3390/rs14030592.
Sabzi, S., Abbaspour-Gilandeh, Y., Arribas, J.I., 2020. An automatic visible-range video weed detection, segmentation and classification prototype in potato field. Heliyon 6. doi:10.1016/j.heliyon.2020.e03685.
Shorewala, S., Ashfaque, A., Sidharth, R., Verma, U., 2021. Weed density and distribution estimation for precision agriculture using semi-supervised learning. IEEE Access 9, 27971–27986. doi:10.1109/ACCESS.2021.3057912.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition URL: http://arxiv.org/abs/1409.1556.
Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C., Cubuk, E.D., Kurakin, A., Li, C., 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence, in: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual. URL: https://proceedings.neurips.cc/paper/2020/hash/06964dce9addb1c5cb5d6e3d9838f733-Abstract.html.
Sokolova, M., Japkowicz, N., Szpakowicz, S., 2006. Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation, pp. 24–29. doi:10.1007/11941439_114.
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2015. Rethinking the inception architecture for computer vision URL: http://arxiv.org/abs/1512.00567.
Tan, M., Le, Q.V., 2021. Efficientnetv2: Smaller models and faster training URL: http://arxiv.org/abs/2104.00298.
Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results URL: http://arxiv.org/abs/1703.01780.
Tharwat, A., 2018. Classification assessment methods. Applied Computing and Informatics 17, 168–192. doi:10.1016/j.aci.2018.08.003.
Varma, S., Simon, R., 2006. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7. doi:10.1186/1471-2105-7-91.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need URL: http://arxiv.org/abs/1706.03762.
Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H., Yang, R., 2022. Salient object detection in the deep learning era: An in-depth survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 3239–3259. doi:10.1109/TPAMI.2021.3051099.
Wei, L., Zhao, S., Bourahla, O.E.F., Li, X., Wu, F., 2017. Group-wise deep co-saliency detection, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, pp. 3041–3047. URL: https://doi.org/10.24963/ijcai.2017/424, doi:10.24963/ijcai.2017/424.
Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2016. Aggregated residual transformations for deep neural networks

URL: http://arxiv.org/abs/1611.05431.
Yang, X., Song, Z., King, I., Xu, Z., 2021. A survey on deep semi-supervised learning URL: http://arxiv.org/abs/2103.00550.
Zeng, Y., Zhuge, Y., Lu, H., Zhang, L., 2019. Joint learning of saliency detection and weakly supervised semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision.
Zhao, T., Wu, X., 2019. Pyramid feature attention network for saliency detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).