The given task of image segmentation was quite challenging to be faced because the dataset is quite small and classes are highly unbalanced. From these crucial points, we started to develop a strategy to manage properly the dataset.
We decided to use all images from the dataset because our objective is to obtain a model able to generalize well across the different types of photos. Since we noticed that data are equally distributed among the four groups, we selected for the validation set the same percentage of images from each class (stratified sampling). What we wanted to achieve is that the distribution of classes in the validation set reflects the distribution of classes in the test set.
In terms of image preprocessing, we applied a mask able to highlight crop and weed to improve the segmentation around boundaries but, from our tries, we found that there was not an improvement and that's why no image preprocessing is used. Furthermore, we trained the networks with different image resizing noticing that reducing too much the size of the images leads to a decay of performances on IoU metric as expected from our scenario, since very small objects, like weeds, must be classified as well.
Besides, very critical aspects of the training are the choice of a loss function and a proper way to manage the class imbalance. To do that, we investigated different combinations of losses (such as logarithmic IoU loss, dice loss, categorical focal loss, and sparse categorical crossentropy loss) and weighting algorithms (such as Enet, median frequency, and simple frequency weighting). The process we followed was to find for each loss the correct weights choice.

From the model perspective, we started building a fully convolutional model from scratch taking inspiration from [1]. The architecture is symmetric and designed upon Segnet model, using Residual blocks as convolutional blocks to make the network smaller and easier to train while keeping a large receptive field. This first model, tested with different depth and hyperparameters, performed pretty well compared to vanilla architectures of Unet or Segnet.

At this point, we applied the techniques of transfer learning and fine-tuning thinking that reusing the convolutional layers of the pre-trained models in the encoder layers could help with our task, affected by data scarcity.
There are several models available for semantic segmentation and the architecture should be chosen properly depending on the use case. In the choices, we took into account the number of training images and the size of the images. The models applied are again inspired by the simple architecture of Unet and Segnet for their simplicity and effectiveness. To select the segmentation model, our first task was to choose an appropriate base network. In particular, we decided to try different networks: VGG-16, ResNet50, and EfficientNetB7. From our results, the best performing models were based on Unet architecture, using ResNet50 or EfficientNetB7 as encoder.

Finally, investigating the performance improvement due to the increase of image dimensions we realized that to overcome the constraint on computing resources we should resort to the technique of tiling.

For the training, analyzing how the image's tiles content changes with the tile dimension, we decided to resort to tiles of 512x512 with a stride of 512x512, in this way we intended to exploit the receptive field of deep networks such as the ones we used. Indeed, smaller tiles gave us lower performances. Moreover, we considered only patches for which the ground truth mask contained a crop or weed pixel aiming at pushing the learning towards less frequent classes, indeed in this way we do not consider the tiles with only background.
At inference time, there might be a high error on predictions made near the tile's borders, hence we predicted on overlapping patches, i.e. stride smaller than patches dimension, and recombined them to prevent jagged predictions.

References

[1] Andres Milioto, Philipp Lottes, and Cyrill Stachniss. "Real-Time Semantic Segmentation of Crop and Weed for Precision Agriculture Robots Leveraging Background Knowledge in CNNs". In: (May 2018), pp. 2229–2235. DOI: 10.1109/ICRA.2018.8460962.