Journal of Universal Computer Science, vol. 29, no. 5 (2023), 419-431
submitted: 11/6/2022, accepted: 19/1/2023, appeared: 28/5/2023

Semi-Supervised Semantic Segmentation for Identification of Irrelevant Objects in a Waste Recycling Plant

César Domínguez
(Universidad de La Rioja, Logroño, La Rioja, Spain
https://orcid.org/0000-0002-2081-7523, cesar.dominguez@unirioja.es)

Jónathan Heras
(Universidad de La Rioja, Logroño, La Rioja, Spain
https://orcid.org/0000-0003-4775-1306, jonathan.heras@unirioja.es)

Eloy Mata
(Universidad de La Rioja, Logroño, La Rioja, Spain
https://orcid.org/0000-0003-0538-4579, eloy.mata@unirioja.es)

Vico Pascual
(Universidad de La Rioja, Logroño, La Rioja, Spain
https://orcid.org/0000-0003-3576-0889, mvico@unirioja.es)

Lucas Fernández-Cedrón
(SpectralGeo, Logroño, La Rioja, Spain
lucas@spectralgeo.es)

Marcos Martínez-Lanchares
(SpectralGeo, Logroño, La Rioja, Spain
marcos@spectralgeo.es)

Jon Pellejero-Espinosa
(SpectralGeo, Logroño, La Rioja, Spain
jon@spectralgeo.es)

Antonio Rubio-Loscertales
(SpectralGeo, Logroño, La Rioja, Spain
antonio@spectralgeo.es)

Carlos Tarragona-Pérez
(SpectralGeo, Logroño, La Rioja, Spain
carlos@spectralgeo.es)

Abstract: In waste recycling plants, measuring the waste volume and weight at the beginning of the treatment process is key for a better management of resources. This task can be conducted by using orthophoto images, but it is necessary to remove from those images the objects that are not involved in the measurement process such as containers or trucks. This work proposes the application of deep learning for the semantic segmentation of those irrelevant objects. Several deep architectures are trained and compared, while three semi-supervised learning methods (PseudoLabeling, Distillation and Ensemble Distillation) are proposed to take advantage of non-annotated

images. In these experiments, the U-net++ architecture with an EfficientNetB3 backbone, trained with the set of labelled images, achieves the best overall multi Dice score of 91.48%. The application of semi-supervised learning methods further boosts the segmentation accuracy in a range between 1.82% and 3.92%, on average.

Keywords: Waste management, Deep Learning, Semi-Supervised Learning, Semantic Segmentation, Orthophoto
Categories: I.2, I.4
DOI: 10.3897/jucs.87643

1 Introduction

As nations and cities become more populated and prosperous, offer more products and services to citizens, and participate in global trade and exchange, they face corresponding amounts of waste to manage through treatment and disposal. By 2050, the world is expected to generate 3.40 billion tons of waste annually, increasing drastically from today's 2.01 billion tons [Kaza et al., 2018]. Therefore, efficient recycling strategies are critical to reduce the devastating environmental effects of rising waste production [Bashkirova et al., 2021]. In this context, waste recycling plants are central since, in these plants, the collected recyclable waste is sorted into separate bales of plastic, paper, metal and glass.

In order to achieve a better management of resources in waste recycling plants, a key indicator is the waste volume and weight at the beginning of the treatment process. This measurement can be performed by using orthophoto images (that is, aerial images that have been geometrically corrected such that their scale is uniform and true distances can be measured) [Ortenzi et al., 2021]; however, it is necessary to process those images to identify and discard objects (like containers or trucks) that might appear in the image, but should not be taken into account in the measurement process — this task is known as object removal, and plays a key role as a pre-processing step to measure properties of objects in images [Pally and Samadi, 2022]. This issue can be faced by means of semantic segmentation algorithms that serve to classify every pixel of an image among target classes of interest [Gonzalez et al., 2002]. Currently, semantic segmentation tasks are mainly tackled by using deep learning methods [LeCun et al., 2015].

Deep learning has many applications in waste management including waste classification [Huang et al., 2020, Meng and Chu, 2020], waste object localisation in outdoor scenarios [Sousa et al., 2019, Proença and Simões, 2020], waste detection and segmentation in materials recovery facilities [Bashkirova et al., 2021], or recognising composition of construction waste mixtures [Lu et al., 2022]. Deep learning methods have been recently used for classifying and detecting objects in waste recycling facilities. For instance, in [Yang et al., 2022], the YOLO object detection algorithm was used to detect electrical and electronic equipment that contain lithium batteries since they have to be processed differently on waste disposal plants; and, two classifiers were combined to classify recyclables and distinguish types of plastics in [Vogiatzis et al., 2021]. There are also a few works that deal with semantic segmentation tasks in waste recycling facilities. Namely, [Bchir et al., 2021] employed a DeepLabv3+ model to segment Polyethylene Terephthalate objects automatically; and, [Sievert, 2021] conducted a comparison of instance segmentation models to guide unmanned vehicles for autonomous litter collection. Finally, the Visual Domain Adaptation 2022 Challenge was recently released with the aim of developing models for automatically industrial waste sorting [Bashkirova et al., 2022].

However, and up to the best of our knowledge, deep learning methods have not been used for segmenting objects in actual waste recycling plants. One of the main challenges for successfully applying deep learning methods is the necessity of a great amount of images that must be manually annotated. Such an annotation process is a tedious and time-consuming task that can take several hours or even days [Lin et al., 2014, Li et al., 2020]. In order to reduce such a burden, close-transfer learning [Razavian et al., 2014] and semi-supervised methods [Zhu and Goldberg, 2009] can be applied. The former methods use the knowledge learned on a close task where acquiring images in easier than in the final task; whereas, the latter methods take advantage of both labelled and unlabelled data. These two approaches have been studied in this work. Namely, we are focused on combining close-transfer techniques and semi-supervised learning methods with deep learning models to produce the exact segmentation of objects that appear in orthophotos of recycling plants, but that should be removed to precisely measure waste volumes. The original contribution of this paper is threefold:

- The analysis of a close-domain transfer learning approach and three semi-supervised learning models to deal with the small size of the annotated dataset by taking advantage of raw images and unlabelled orthophoto images.
- A detailed comparison of several state-of-the-art deep neural networks for semantic segmentation (both architectures and backbones) for processing orthophoto images.
- A statistical analysis to identify whether there are significant differences among the studied deep learning models and the semi-supervised learning methods.

As a result, this paper demonstrates that using a semi-supervised learning technique allows us to successfully train segmentation models, substantially reducing the effort required to annotate many images. Then, the comparison of the networks shows that the U-net++ architecture with the EfficientNetB3 backbone achieves the best performance in segmentation accuracy. In this case, the multi Dice score is equal to 91.65%. This result confirms that orthophoto images can be effectively processed to segment the environment of recycling plants and find objects of interest. This step will enable the use of orthophoto images for measuring waste volume with a higher precision.

The rest of the paper is organised as follows: the first section on Materials and Methods describes the input datasets, the semantic segmentation models, the semi-supervised learning methods and the way results are evaluated and compared; the Experimental Result section presents the outcomes of the tests; and, the last section ends the manuscript with final comments and remarks on future activities.

2 Materials and methods

2.1 Input dataset

This paper tackles the problem of image segmentation from orthophoto images captured in a recycling plant. Within these lines, the automatic segmentation of orthophoto images is achieved by representing them in more descriptive and discriminative feature spaces, learned from actual images, where pixels having similar semantic attributes can be grouped and labelled in different classes. A set of annotated images is thus required to allow the training of the models. At the same time, additional annotated images are

Figure 1: Sample orthophoto image and corresponding annotated image. black pixels belong to the class person, yellow to the container class, red to the forklift class, green to the truck class, and black to background

needed to evaluate the classification results on a ground truth. These two sets of images (training and test sets) form the annotated dataset.

The annotated dataset is made of 49 manually-labelled colour images acquired by a Safire IP 8MPx camera with a focal of 2.8mm in a recycling plant in Spain. In order to construct the orthophotos, 10 images of resolution 2560×1440 pixels were taken and combined using the ODM technology through the PyODM library [OpenDroneMap project, 2021]. The orthophoto images have a resolution of 1526×1468 pixels, and were manually annotated using the Labelme tool [Wada et al., 2021] as shown in Figure 1. The annotation aims to separate five classes of interest: person (black segments), container (yellow segments), forklift (red segments), truck (green segments), and background (black segments). All the images were taken from the same facility, but there is a considerable variability in the images since they are taken on different days and with different amounts of waste. Namely, the number and position of containers in the images varies from image to image; besides, the amount of waste in them changes from image to image. Similarly, the number of people and their position changes from image to image. Finally, the two classes of objects with most variability are forklift and truck since the generation of the orthophotos deforms those objects in the images.

The manual annotation of images is a time-demanding and tedious task. Although the acquisitions provide many images, the annotation has been limited to the representative subset of 49 images described previously. However, there are 322 further orthophoto images, not-labelled but acquired under the same experimental conditions. These orthophoto images will tune the training of the networks through the implementation of three semi-supervised approaches. The network architectures and the semi-supervised algorithms will be detailed in the following subsection.

2.2 Semantic segmentation models

As stated in the previous section, the 49 labelled orthophoto images, randomly splitted into training sets (39 images) and test sets (10 images) using a 5-fold cross-validation approach, were used to set up and evaluate the deep segmentation architectures (see Table 1 for the number of objects in each dataset). From the training set, several deep-learning segmentation algorithms were fine-tuned [Razavian et al., 2014]. Namely, 7 architectures

	Person	Container	Forklift	Truck
Training set 1	34	296	38	10
Test set 1	8	75	9	2
Training set 2	32	297	38	9
Test set 2	10	74	9	3
Training set 3	32	293	38	11
Test set 3	10	78	9	1
Training set 4	32	292	37	11
Test set 4	10	79	10	1
Training set 5	38	306	37	7
Test set 5	4	65	10	5

Table 1: Number of objects of interest in the training and test sets

were trained, they are summarised in Table 2 — we fixed a seed for reproducibility and train each model just once. For training, we used the libraries PyTorch [Paszke et al., 2019] and FastAI [Howard and Gugger, 2020]; and using a GPU Nvidia RTX 2080 Ti. The procedure presented in [Howard and Gugger, 2020] was employed to set the learning rate for the different architectures, the learning rate for the first layers of the models was fixed to 1e-4, and for the last layers of the models to 1e-3. Also, early stopping was applied in all the architectures to avoid overfitting (validation loss was monitored and the training process stopped when such a validation loss did not improve after 5 epochs). As a result of the training process, several models were produced that can be used for inference by providing them a natural image as input. Then, the models will output the mask associated with the segmentation.

In addition, we have applied a close transfer-learning approach for training our models. When applying transfer learning, it is well-known the importance of using a source task that is as close as possible to the target task [Mensink et al., 2021]. Therefore, we have used 404 raw images that were used to construct the orthophotos of the training set (390 images were used to generate the 39 ortophoto images of the training dataset, and the other 14 raw images were introduced to increase the variability of the dataset, but were not used for generating orthophoto images). These 404 images were manually annotated, and used to train the models from Table 2 — annotating raw images in easier than annotating orthophoto images due to the distortions that might appear in the latter images. Subsequently, those models were used as starting point to train the same models but using the orthophoto images.

Using a 5-fold cross-validation approach, all the models were then evaluated on the test set of 10 annotated orthophoto images using the multi-class Dice score [Opitz and Burst, 2019]. This metric is defined using the precision, P_i, and recall, R_i, values for each class defined as:

$$P_i = \frac{m_{ii}}{\sum_{x=1}^{n} m_{ix}}, \quad R_i = \frac{m_{ii}}{\sum_{x=1}^{n} m_{xi}}$$

where n is the number of classes, and m_{jk} for $j = 1, \ldots, n$ and $k = 1, \ldots, n$ is the total number of pixels predicted as $class_j$, whose actual label is $class_k$. From the precision

Architecture	Backbones
Bisenet	Resnet18, Resnet34
Deeplabv3+	Resnet50, Resnext50, EfficientNetB3
HRNet	w30
Manet	Resnet50, Resnext50, EfficientNetB3
PAN	Resnet50, Resnext50, EfficientNetB3
U-net	Resnet50, Resnest50, EfficientNetB3
U-net++	Resnet50, Resnest50, EfficientNetB3

Table 2: Segmentation architectures and the backbones employed in this work

and recall values, the multi-class Dice score is defined as follows:

$$MultiDice = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i}$$

2.3 Semi-supervised learning methods

In order to take advantage of the unlabelled images, 3 semi-supervised learning approaches were employed. Namely, we have employed PseudoLabeling [Lee, 2013], Distillation [Hinton et al., 2015], and Ensemble Distillation [Bucila et al., 2006] — the latter method is also known as Model Distillation.

The PseudoLabeling approach consists of two steps; first, we employ a model trained on a manually labelled dataset to make predictions in an unlabelled dataset; secondly, the manually and automatically-labelled datasets are combined to train a new model using the same architecture employed in the original model. We have applied the PseudoLabeling approach to all the architectures presented in the previous section; and, the initial model was trained with the close-transfer learning approach.

The Distillation approach is similar to the PseudoLabeling approach, but in the second step, the model might have a different underlying architecture than the model employed for the first step. In our case, we have trained several models using the training procedure presented in the previous section, and selected the best model for generating the automatically-labelled dataset. Furthermore, we have used the combination of the manually and automatically-labelled datasets to train all the architectures presented in the previous section.

Finally, Ensemble Distillation differs from the Distillation approach in the way of producing the automatically labelled dataset; namely, instead of using a single model for making predictions in an unlabelled dataset, the predictions are generated from an ensemble of models. In this work, we have employed the 5 models with the highest total multi Dice score for producing the predictions on the unlabelled dataset; and, as in the previous approaches, the manually and automatically-labelled datasets were used to train all the architectures presented in the previous section.

3 Experimental Results

The performance of the trained networks (both by applying and without applying the semi-supervised learning methods) was evaluated using a 5-fold cross-validation approach

Model	Baseline	Close-Transfer	Pseudolabel	Distillation	Ensemble distillation
Bisenet-Resnet18	52.07 (11.82)	87.52 (0.36)	87.48 (0.47)	88.41 (0.4)	88.39 (0.12)
Bisenet-Resnet34	51.74 (13.58)	86.91 (0.33)	86.65 (0.72)	87.68 (0.4)	88.12 (0.34)
Deeplabv3+-Resnet50	72.06 (6.83)	90.05 (0.07)	89.74 (0.23)	90.06 (0.37)	90.0 (0.3)
Deeplabv3+-Resnext50	74.82 (6.9)	89.0 (0.16)	89.73 (0.23)	90.27 (0.16)	89.54 (0.17)
Deeplabv3+-EfficientNetB3	73.15 (5.84)	90.27 (0.2)	90.17 (0.33)	90.47 (0.17)	90.37 (0.46)
HRNet-w30	48.89 (13.68)	85.51 (6.81)	84.23 (7.25)	90.64 (0.54)	90.58 (0.18)
Manet-Resnet50	63.28 (12.59)	66.33 (5.66)	68.81 (1.08)	81.69 (9.52)	81.31 (9.32)
Manet-Resnest50	57.01 (11.27)	80.06 (7.93)	89.58 (0.5)	87.26 (7.68)	90.5 (0.72)
Manet-EfficientNetB3	68.06 (11.35)	56.17 (0.45)	55.48 (16.63)	64.38 (6.32)	59.08 (4.46)
PAN-Resnet50	66.87 (9.38)	72.24 (6.23)	76.0 (7.47)	76.83 (8.53)	64.1 (3.99)
PAN-Resnext50	64.84 (17.86)	48.43 (0.94)	54.33 (4.65)	60.62 (5.78)	62.68 (8.43)
PAN-EfficientNetB3	68.83 (7.14)	68.06 (0.44)	81.13 (0.01)	82.64 (7.18)	85.57 (1.03)
U-net-Resnet50	70.11 (0.59)	89.22 (0.75)	89.85 (0.25)	90.65 (0.16)	90.63 (0.53)
U-net-Resnest50	72.27 (3.83)	86.82 (1.39)	85.31 (1.11)	83.19 (8.67)	89.95 (0.44)
U-net-EfficientNetB3	73.58 (5.62)	90.66 (0.29)	90.68 (0.16)	91.03 (0.18)	91.06 (0.4)
U-net++-Resnet50	73.89 (6.13)	89.8 (0.48)	90.28 (0.18)	90.64 (0.2)	90.73 (0.19)
U-net++-Resnest50	74.15 (6.88)	90.36 (0.14)	90.54 (0.08)	91.11 (0.29)	90.73 (0.21)
U-net++-EfficientNetB3	79.12 (6.87)	90.93 (0.27)	91.2 (0.16)	91.35 (0.15)	**91.48 (0.3)**

Table 3: Mean (std) multi Dice score from the application of the different learning procedures. The result in bold is the best

with independent test sets that consist of 10 orthophoto images.

All the models trained only with the manually labelled orthophoto images (from now on, we will refer to this approach as baseline) achieved a multi Dice score under 80%, and only the model trained with the U-net++ architecture and the EfficientNetB3 backbone obtained a multi Dice score over 75%, see Table 3.

The results can be considerably improved by applying the close-transfer learning approach, see Table 3. Namely, some models improved more than a 35%, and on average a 14.09%. If the segmentation networks are compared, there are five networks (Deeplabv3+-Resnet50, Deeplabv3+-EfficientNetB3, U-net-EfficientNetB3, U-net++-Resnest50 and U-net++-EfficientNetB3) with a total multi Dice score over 90%. Among them, the U-net++ architecture together with the EfficientNetB3 backbone showed better segmentation accuracy than the other networks. Namely, this model achieved a multi Dice score of 90.93%, and the improvement regarding its baseline counterpart is shown in Figure 2.

The impact of the different semi-supervised learning methods for the studied networks is also shown in Table 3. With more details, the Distillation approach produces a mean improvement of 3.92% (with a standard deviation of 5.46%). Only one network get worse results using this training approach while, in some cases, namely for the Manet-Resnet50 model, the improvement is over 15%, see Figure 3. Similarly, the PseudoLabeling method produces a mean improvement of 1.82% (with a standard deviation of 3.93%), with seven networks having worse results. Finally, the Ensemble Distillation method also considerably improves the performance of the models (a mean of 3.69% with a standard deviation of 6.48%).

In our analysis, we can see that there are some architectures that generally perform better than the others independently of the training method. The architectures that obtained worse results were two architectures based on the attention mechanism: PAN [Li et al., 2018] and Manet [Li et al., 2021]. This kind of network requires more images to be

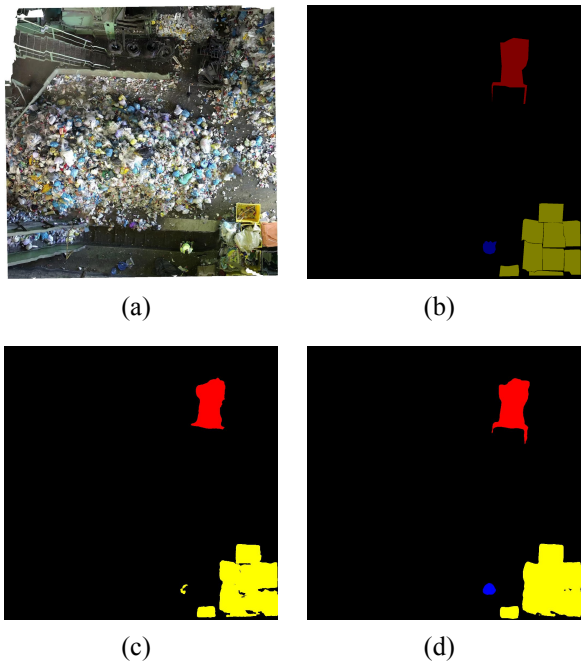

Figure 2: Example of the segmentation results using the U-net++-EfficientNetB3 model trained with the baseline and close-transfer approaches. (a) Input image, (b) Ground truth, (c) Baseline model, (d) Close-transfer model

trained properly [Dosovitskiy et al., 2021], and unfortunately we are working with a small dataset. The other architectures are based on convolutional neural networks, and obtained better results; however, we can also notice some differences among them. The Bisenet models obtained worse results due to the size of their backbones (Resnet18 and Resnet34) that are smaller than other backbones such as Resnet50 or EfficientNetB3. In the case of the U-net models, the U-net++ version achieved a higher performance due to the several improvements introduced in this new version of the U-net architecture [Zhou et al., 2018]. Finally, the DeepLabv3, HRNet, and U-net++ obtained similar results independently of the selected backbone.

In addition to searching for the best performing model, we have conducted a statistical study to determine whether the results obtained with the different training approaches are statistically significant. To this aim, several null hypothesis tests have been performed using the methodology presented in [Garcia et al., 2010, Sheskin, 2011]. In order to choose between a parametric or a non-parametric test to compare the models, we check three conditions: independence, normality and heteroscedasticity — the use of a parametric test is only appropriate when the three conditions are satisfied [Garcia et al., 2010].

In this study, the independence condition is fulfilled since each semi-supervised learning approach is independent of the others. We use the Shapiro-Wilk test [Shapiron and Wilk, 1965] to check normality — with the null hypothesis being that the data follow a normal distribution — and, a Levene test [Levene, 1960] to check heteroscedasticity — with the null hypothesis being that the results are heteroscedastic. Since we compare

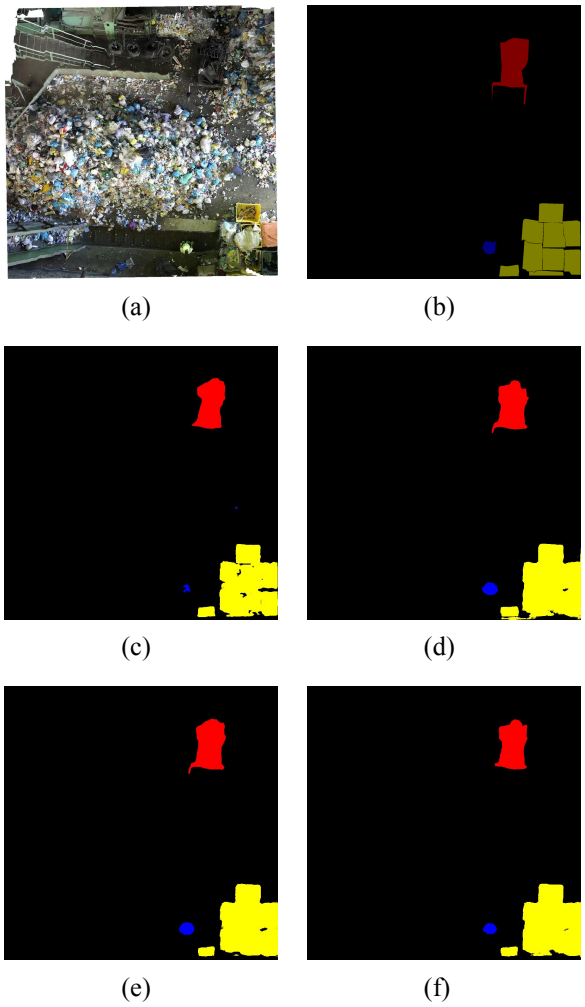

Figure 3: Example of the segmentation results using the semi-supervised learning approaches with the U-net-Resnet50 model. (a) Input image, (b) Ground truth, (c) Close-transfer model, (d) Distillation, (e) PseudoLabelling, (f) Ensemble Distillation

more than two training approaches, we will employ an ANOVA test if the parametric conditions are fulfilled, and a Friedman test otherwise [Sheskin, 2011]. In both cases, the null hypothesis will be that all the models have the same performance. Once the test for checking whether a model is statistically better than the others is conducted, a post-hoc procedure is employed to address the multiple hypothesis testing among the different models. A Holm post-hoc procedure [Holm, 1979], in the non-parametric case, or a Bonferroni-Dunn post-hoc procedure [Sheskin, 2011], in the parametric case, is used for detecting significance of the multiple comparisons [Garcia et al., 2010, Sheskin, 2011] and the p values should be corrected and adjusted. We have performed our experimental

analysis with a level of confidence equal to 0.05. In addition, the size effect has been measured using Cohen's d [Cohen, 1969] and Eta Squared [Cohen, 1973].

In our study, and since the normality condition was not fulfilled (Shapiro-Wilk's test W=0.818257; $p < 0.001$), a Friedman's non-parametric test was employed to compare the training procedures. Friedman's test performed a ranking of the training procedures under comparison (see Table 4), assuming as null hypothesis that all the models have the same performance. In this case, significant differences arise ($F = 19.05$; $p = 1.51e^{-10}$) with a large size effect Eta Squared 0.29. The Distillation method produced the best models.

Technique	Dice score	Friedman's test average ranking
Distillation	84.94 (8.89)	4.28
Ensemble distillation	84.71 (10.48)	4.06
Pseudolabel	82.84 (11.40)	2.72
Close-Transfer	81.02 (12.75)	2.39
Baseline	66.96 (8.68)	1.55

Table 4: Friedman's test for the Dice score of the segmentation models

In Table 5, we show the results of the application of the Holm post-hoc procedure to compare the control training procedure (winner, based on distillation) with all the other training approaches, adjusting the p-value. Results prove significant differences between the semi-supervised learning procedures and the plain training approaches, while all the semi-supervised learning methods produce similar outcomes. The size effect is also taken into account using Cohen's d, and, as shown in Table 5; and, it is large size when the winning approach is compared with the plain training approach.

Technique	Z value	p value	adjusted p value	Cohen's d
Baseline	5.16	2.40e-07	9.61e-07	1.99
Close-Transfer	3.58	0.0003	0.0010	0.35
Pseudolabel	2.95	0.0031	0.0063	0.20
Ensemble distillation	0.42	0.6733	0.6733	0.02

Table 5: Adjusted p-values with Holm and Cohen's d. Control technique: Distillation

4 Conclusion and further work

In this work, we have presented an approach to identify irrelevant objects in a waste recycling plant from ortophoto images. Our method was based on training several deep

learning models for segmenting those irrelevant objects — this approach achieved a multi Dice score of approximatelly 80%. The main limitation of such an approach was the reduced number of available annotated ortophoto images. This drawback was tackled using two different approaches: a close transfer learning method, and several semi-supervised learning techniques. The former used semantic segmentation models pre-trained on raw images, that are easier to annotate than orthophoto images. Thanks to that pre-training stage, the performance of our model improved up to a 90.93%. Finally, the application of semi-supervised learning methods further boosted multi Dice score in a range between 1.82% and 3.92%, on average. Therefore, we have presented several methods to improve segmentation accuracy by taking advantage of raw images and unlabelled orthophoto images, thus avoiding the need for a large dataset of labelled images, whose annotation can be time-demanding.

For further work, we plan to use the developed models to remove the irrelevant objects from the images, and then use those images to predict the waste volume and weight at the beginning of the treatment process in waste recycling plants.

Compliance with Ethical Standards

Conflict of Interest: All the authors declare that they have no conflict of interest.
Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

Acknowledgements

This work was partially supported by Ministerio de Ciencia e Innovación [PID2020-115225RB-I00 / AEI / 10.13039/501100011033], and the SEPARA project funded by a CDTI project of the Misiones 2020 call.

References

[Bashkirova et al., 2021] Bashkirova, D. et al. (2021). Zerowaste dataset: Towards deformable object segmentation in cluttered scenes. arXiv preprint, abs/2106.02740.

[Bashkirova et al., 2022] Bashkirova, D. et al. (2022). Visual domain adaptation challenge (visda-2022). https://ai.bu.edu/visda-2022/.

[Bchir et al., 2021] Bchir, O., Alghannam, S., Alsadhan, N., Alsumairy, R., Albelahid, R., and Almotlaq, M. (2021). Computer vision based polyethylene terephthalate (pet) sorting for waste recycling. *International Journal of Advanced Computer Science and Applications*, 12(10).

[Bucila et al., 2006] Bucila, C., Caruana, R., and Niculescu-Mizil, A. (2006). Model compression: making big, slow models practical. In *Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining*, KDD'06, pages 535–541.

[Cohen, 1969] Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, USA.

[Cohen, 1973] Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor anova designs. *Educational and Psychological Measurement*, 33:107–112.

[Dosovitskiy et al., 2021] Dosovitskiy, A. et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *2021 International Conference on Learning Representations (ICLR)*.

[Garcia et al., 2010] Garcia, S. et al. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences*, 180:2044–2064.

[Gonzalez et al., 2002] Gonzalez, R. C., Woods, R. E., et al. (2002). *Digital image processing*. Prentice hall Upper Saddle River, NJ.

[Hinton et al., 2015] Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the Knowledge in a Neural Network. *CoRR*, abs/1503.02531.

[Holm, 1979] Holm, O. S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70.

[Howard and Gugger, 2020] Howard, J. and Gugger, S. (2020). Fastai: A layered api for deep learning. *Information*, 11:108.

[Huang et al., 2020] Huang, G.-L., He, J., Xu, Z., and Huang, G. (2020). A combination model based on transfer learning for waste classification. *Concurrency and Computation: Practice and Experience*, 32(19):e5751.

[Kaza et al., 2018] Kaza, S., Yao, L. C., Bhada-Tata, P., and Woerden, F. V. (2018). *What a Waste 2.0 : A Global Snapshot of Solid Waste Management to 2050*. World Bank, Washinton DC, USA.

[LeCun et al., 2015] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

[Lee, 2013] Lee, D. H. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Proceedings ICML Workshop: Challenges in Representation Learning (WREPL)*.

[Levene, 1960] Levene, H. (1960). *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, chapter Robust tests for equality of variances, pages 278–292. Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. Stanford University Press, USA.

[Li et al., 2020] Li, G., Wan, J., He, S., Liu, Q., and Ma, B. (2020). Semi-supervised semantic segmentation using adversarial learning for pavement crack detection. *IEEE Access*, 8:51446–51459.

[Li et al., 2018] Li, H., Xiong, P., An, J., and Wang, L. (2018). Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*.

[Li et al., 2021] Li, R., Zheng, S., Zhang, C., Duan, C., Su, J., Wang, L., and Atkinson, P. M. (2021). Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13.

[Lin et al., 2014] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

[Lu et al., 2022] Lu, W., Chen, J., and Xue, F. (2022). Using computer vision to recognize composition of construction waste mixtures: A semantic segmentation approach. *Resources, Conservation and Recycling*, 178:106022.

[Meng and Chu, 2020] Meng, S. and Chu, W.-T. (2020). A study of garbage classification with convolutional neural networks. In *2020 Indo–Taiwan 2nd International Conference on Computing, Analytics and Networks (Indo-Taiwan ICAN)*, pages 152–157. IEEE.

[Mensink et al., 2021] Mensink, T., Uijlings, J., Kuznetsova, A., Gygli, M., and Ferrari, V. (2021). Factors of influence for transfer learning across diverse appearance domains and task types. *arXiv preprint arXiv:2103.13318*.

[OpenDroneMap project, 2021] OpenDroneMap project (2021). Pyodm: a library for easily creating orthophotos. https://pyodm.readthedocs.io/en/latest/.

[Opitz and Burst, 2019] Opitz, J. and Burst, S. (2019). Macro f1 and macro f1. *arXiv preprint arXiv:1911.03347*.

[Ortenzi et al., 2021] Ortenzi, L., Violino, S., Pallottino, F., Figorilli, S., Vasta, S., Tocci, F., Antonucci, F., Imperi, G., and Costa, C. (2021). Early estimation of olive production from light drone orthophoto, through canopy radius. *Drones*, 5(4).

[Pally and Samadi, 2022] Pally, R. and Samadi, S. (2022). Application of image processing and convolutional neural networks for flood image classification and semantic segmentation. *Environmental Modelling & Software*, 148:105285.

[Paszke et al., 2019] Paszke, A. et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

[Proença and Simões, 2020] Proença, P. F. and Simões, P. (2020). Taco: Trash annotations in context for litter detection. *arXiv preprint arXiv:2003.06975*.

[Razavian et al., 2014] Razavian, A. S., Azizpour, H., Sullivan, J., et al. (2014). CNN features off-the-shelf: An astounding baseline for recognition. In *CVPRW'14*, pages 512–519.

[Shapiron and Wilk, 1965] Shapiron, S. S. and Wilk, M. B. (1965). An analysis for variance test for normality (complete samples). *Information Sciences*, 180:2044–2064.

[Sheskin, 2011] Sheskin, D. (2011). *Handbook of Parametric and Nonparametric Statistical Procedures*. CRC Press, London.

[Sievert, 2021] Sievert, R. (2021). Instance segmentation of multiclass litter and imbalanced dataset handling: A deep learning model comparison.

[Sousa et al., 2019] Sousa, J., Rebelo, A., and Cardoso, J. S. (2019). Automation of waste sorting with deep learning. In *2019 XV Workshop de Visão Computacional (WVC)*, pages 43–48. IEEE.

[Vogiatzis et al., 2021] Vogiatzis, A., Chalkiadakis, G., Moirogiorgou, K., Livanos, G., Papadogiorgaki, M., and Zervakis, M. (2021). Dual-branch cnn for the identification of recyclable materials. In *2021 IEEE International Conference on Imaging Systems and Techniques (IST)*, pages 1–6. IEEE.

[Wada et al., 2021] Wada, K. et al. (2021). Labelme: image polygonal annotation with python. https://github.com/wkentaro/labelme.

[Yang et al., 2022] Yang, S. W., Park, H. J., Kim, J. S., Choi, W., Park, J., and Han, S. W. (2022). Study on the real-time object detection approach for lithium-based secondary battery in WEEE recycling process. *Available at SSRN 4181525*.

[Zhou et al., 2018] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer.

[Zhu and Goldberg, 2009] Zhu, X. and Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130.