Synthesized Image Datasets: Towards an Annotation-Free Instance Segmentation Strategy

Henry O. Velesaca[1], Patricia L. Suárez[1], Dario Carpio[1], and Angel D. Sappa[1,2](✉)

[1] ESPOL Polytechnic University, FIEC, CIDIS, Guayaquil, Ecuador
{hvelesac,plsuarez,dncarpio,asappa}@espol.edu.ec
[2] Computer Vision Center, 08193 Bellaterra, Barcelona, Spain
asappa@cvc.uab.es

Abstract. This paper presents a complete pipeline to perform deep learning-based instance segmentation of different types of grains (e.g., corn, sunflower, soybeans, lentils, chickpeas, mote, and beans). The proposed approach consists of using synthesized image datasets for the training process, which are easily generated according to the category of the instance to be segmented. The synthesized imaging process allows generating a large set of well-annotated grain samples with high variability—as large and high as the user requires. Instance segmentation is performed through a popular deep learning based approach, the Mask R-CNN architecture, but any learning-based instance segmentation approach can be considered. Results obtained by the proposed pipeline show that the strategy of using synthesized image datasets for training instance segmentation helps to avoid the time-consuming image annotation stage, as well as to achieve higher intersection over union and average precision performances. Results obtained with different varieties of grains are shown, as well as comparisons with manually annotated images, showing both the simplicity of the process and the improvements in the performance.

Keywords: Instance segmentation · Food grains · Synthesized dataset generation

1 Introduction

Deep learning based approaches have shown to be the best option to tackle challenging computer vision problems such as segmentation [14], recognition [13,15], 3D estimation [21], scene understanding [23] just to mention a few. Actually, deep learning based solutions have become the facto approaches to tackle challenging tasks in computer vision, as well as in many fields. Although there are deep learning based models able to obtain good results with a few training samples, in most of the cases their performance depends on the amount of annotated data available during the training process.

Most of deep learning based models, for instance ResNet [8], VGG [16], AlexNet [10], Mask R-CNN [7], are able to reach very accurate results for different applications when trained on large and well-annotated datasets. Unfortunately, collecting annotations at scale is not feasible or it is prohibitively expensive. Actually, this expensive

G. Bebis et al. (Eds.): ISVC 2021, LNCS 13017, pp. 131–143, 2021.
https://doi.org/10.1007/978-3-030-90439-5_11

task has been tackled in recent year by different initiatives for a few number of categories (e.g., pedestrians, cars, dogs, trains, among others) resulting in the well known PASCAL VOC [5], COCO [12], ImageNet [4], SUN [22] datasets. These datasets contain hundred of thousand of images with millions of annotations (bounding boxes, used for recognition tasks), or thousands of well-annotated masks (instance's contour, used for segmentation applications).

The datasets mentioned above have been the starting point to develop interesting and useful applications for the video surveillance [6] or driving assistance [20] fields, where categories such as a person, car, bike, among others, are needed for training object detection algorithms, or regions correctly annotated in urban scenarios for semantic segmentation tasks. A bottleneck of most of deep learning based approaches lies is the need of having large and well-annotated instances. In the current work, the seed segmentation problem is tackled. In other words, given a cluster of crowded instances, the algorithm should return the boundary of every single instance in the scene. Although there are robust and efficient architectures to solve this problem (e.g., Mask R-CNN [7], YOLACT [2], Deep watershed transform [1]), their performance is highly affected by the dataset used for training; not only the quantity of instances in the given datasets is important, but also the quality of annotations (i.e., objects' boundary) is a key factor. Furthermore, it is not easy to find datasets that adapt to the requirements of different tasks.

Having in mind the limitations mentioned above, the current work proposes a novel strategy to generate annotated images to be used for training instance segmentation algorithms. Although the proposed strategy is evaluated on the well-known Mask R-CNN [7], it can be also applied with other instance segmentation models. The main contribution lies in the pipeline that allows to automatically generate annotated synthesized images that can be used for training or to extend manually annotated datasets. This mainly reduces time and annotation effort and allows the network to easily be trained for different scenarios and acquisition sensors. The segmentation of instances of different types of grains is approached, taking into consideration the corn grains as a case study to evaluate the effectiveness of the use of synthesized images comparing them with the results obtained using only real images.

The manuscript is organized as follows. Section 2 presents works related to the instance segmentation as well as recent approaches on grain segmentation. The approach proposed for generating synthesized datasets is introduced in Sect. 3. Experimental results with different categories and evaluations using ground truth images are depicted in Sect. 4. Finally, conclusions are presented in Sect. 5.

2 Related Work

As mentioned above, current work addresses the problem of kernel instance segmentation and the need for datasets of large size and variability. In other words, this work is focused on obtaining all the instances (i.e., grains) present in a cluster image with a random distribution. This section reviews the most relevant works on these topics, highlighting the main characteristics of each of the reviewed approaches.

Regarding instance segmentation, the Mask R-CNN architecture has become a referent in recent years in the area of object detection and instance segmentation; it extends

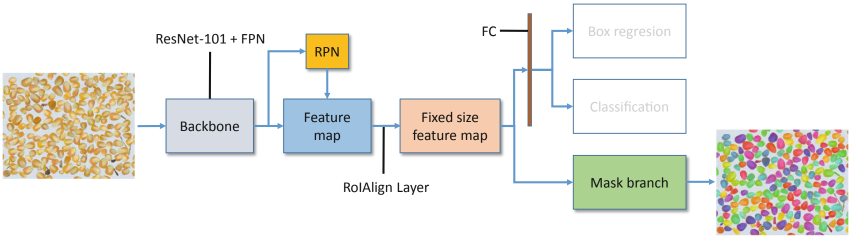

Fig. 1. Mask R-CNN architecture used for grain instance segmentation. Images of synthesized clusters of corn grains are shown as an example (classification module is not used).

the Faster R-CNN object detection framework by adding a branch for the generation of the masks at the end of the model, thus achieving the instances segmentation for each output proposal box. In addition, the segmentation is executed in parallel to the identification and location. The Mask R-CNN framework consists of three stages (see Fig. 1). First, the backbone extracts feature maps from the input images. Second, the feature maps generated by the backbone are sent to the Region Proposal Network (RPN) to generate Regions of Interest (RoIs). Third, the ROIs generated by RPN are assigned to extract the corresponding target features in the shared feature maps, and subsequently mapping to a fully connected layer, for target classification and segmentation instances. The process generates the classification scores, bounding boxes, and segmentation masks [7,24].

Recently, some works have been proposed for the instance segmentation of grains. For instance, Toda et al. [17] present an instance segmentation neural network to determine the morphological phenotype of barley grains. The authors propose to use a synthetically generated dataset for the training stage, where the seeds are randomly oriented to give the corresponding variability. The model trained with synthesized images has given better results compared to the training with real images. In addition, the authors validate the strategy in other types of grains such as wheat, rice, oat, and lettuce grains. The proposed strategy allows generating appealing results. Similarly, our approach tackles the generation of synthesized corn kernel clusters by randomly distributing kernels but in our case, the HSV color space is used to obtain the area of every single instance (i.e., corn kernel). In this color space, more precise contours are obtained and shadows are easily removed.

On the other hand, Kar et al. [9] present a system for automatically estimating the quality of food grains in which wheat grains are presented as a case study. In this case, grains are segmented and classified into eight categories to analyze their quality. To carry out this objective, a convolutional network is trained with a dataset that consists of around 5000 synthesized images, according to the authors this model presents good results both at the time of instance segmentation and classification. One of the differences with the proposal in the current work is that the work presented by Kar et al. [9] uses the U-Net network for segmenting single grains which are later on used to generating the synthesized images; in the current work single grains' mask are obtained by simple thresholding in the HSV color space, which consumes much fewer resources and processing time. Another difference from the approach proposed by Kar et al. [9] in which the input grains are randomly distributed concerning ours is that the distribution in grid format provides better robustness because there is no error when segmenting the instances that are used to generate the final synthesized cluster image.

Another works to carry out segmentation of instances before the classification of types and defects in corn grains is the one presented by Velesaca et al. [18]. In that work, the segmentation is performed using the Mask R-CNN network for subsequent classification with a lightweight convolutional network. The segmentation algorithm is trained using a real dataset manually annotated; a crowdsourcing platform, Labelbox[1], has been used to label every single element (e.g., impurities and grains) in the image. This annotation process requires a lot of time and resources.

In [3] the authors present an evaluation of the effectiveness of segmentation using images with uniform and textured regions in synthetically generated gray levels, using unsupervised evaluation criteria based on image regions. Another segmentation technique based on multichannel texture filtering, which according to the authors allows detecting similar regions and determining abrupt changes in patterns at the texture level of the images is presented in [11]. This method was tested using both real and synthetic images of simple patterns and wood grains, obtaining results similar in segmentation by regions. Following the segmentation line, Wang et al. [19] presents a technique that allows segmenting granular rock based on an extension of the skeleton of the image, differentially eroding it to detect the cores of the granular rock. The results show good effectiveness when compared to methods such as watershed.

3 Proposed Approach

This section first summarizes the Mask R-CNN instance segmentation network used in the current work. Then, the proposed strategy for the generation of synthesized images is presented. The proposed strategy has been evaluated through seven different categories of grains: corn, beans, lentils, mote, soybeans, chickpeas, and sunflower. The corn kernel is used as a case study to illustrate the performance of the proposed pipeline; in this case, study instances are manually annotated to be used as ground truth.

3.1 Segmentation Algorithm

This section introduces the instance segmentation algorithm used to validate the synthesized imaging strategy presented in the current work. Among the different approaches, the Mask R-CNN has been selected due to its great performance in the instances segmentation task for different categories. Another characteristic of the Mask R-CNN is that it uses the ResNet 101 architecture to extract features from the image and has a large number of parameters (i.e., 63738 K), which makes it a complex architecture that requires datasets with a lot of images for the training stage.

As mentioned above, the Mask R-CNN network is used for segmenting the given image in instances—i.e., grains present on it. Initially, the model trained with just the

[1] labelbox.com.

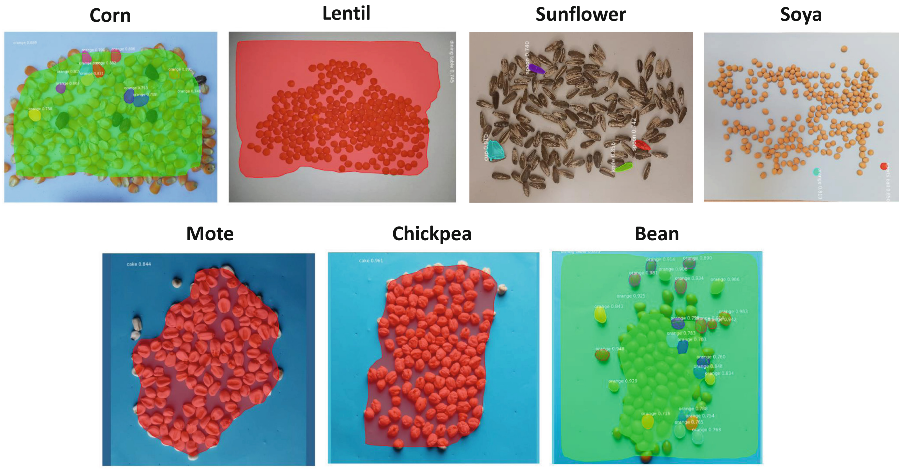

Fig. 2. Examples on different types of grains from the Mask R-CNN network pre-trained with COCO dataset.

COCO dataset is considered—in the 91 classes of COCO there are not seeds, grains, or cereals categories, hence the network is not able to correctly segment the given images. Figure 2 shows illustrations of the results obtained for seven different types of grains; as can be appreciated, inconsistent results are obtained both in the area of each segmented grain and in the number of predicted classes, in some cases (e.g., lentil) the result is just a single big patch. As a conclusion, it can be stated that it is necessary to apply a training process for this particular problem in order to achieve acceptable results in these scenarios. For the training process, a large dataset of annotated instances is required; the image annotation can be performed manually, as performed in [18], or by using the strategy proposed in the current work, which consists of generating synthesized images, which are directly annotated, from single real grains. The synthesized image generation process is presented in the next section.

3.2 Synthesized Image Generation

As mentioned above, training instance segmentation algorithms require a large and well-annotated dataset. This section presents a strategy for avoiding this time-consuming task, as well as for a cost-effective at scale dataset annotation. In this way, large datasets for different varieties of grains can be easily generated for the training process. The proposed approach consists of two tasks: firstly images of single grains are acquired from real scenes, and secondly, synthesized images of a cluster of grains are generated by using sets of single grains. The number of grains in the resulting cluster as well as their distribution can be set by the user as detailed below. Figure 3 shows the pipeline proposed to obtain the synthesized images of different types of grains that will be then applied to the case study of corn kernels.

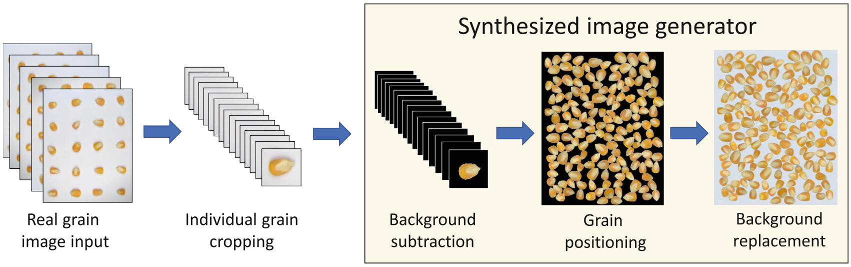

Fig. 3. Overall pipeline for synthesized clusters of grains generation. Firstly, each grain from the real images—i.e., grids of grains—is cropped. Then, the grain's area is extracted using background subtraction. Next, grains are inserted into an empty synthetic image by the positioning algorithm; and finally, a custom background is applied to the synthesized image.

The first task for the generation of a synthesized image dataset consists of acquiring images with samples of non-touching kernels—grids of kernels. The kernel sample acquisition process consists of taking images of an A4 sheet where grains are uniformly distributed. The amount of grains per sheet as well as the color of the sheet depends on the type of grain, for instance, a blue sheet has been used for the mote and chickpea grains, while other colors have been considered in lentil or sunflower to obtain a higher contrast between grains and background. In each case, a grid is drawn to split the image up into small regions containing a single grain in each cell. Images have been acquired with a mobile device (12MP images) orthogonal to the A4 sheet. The resulting ROI images are used in the next step to generate the synthesized images of a cluster of grains. Figure 3$(left)$ shows an illustration of an image with single kernels (i.e., corn grains) uniformly distributed on a light gray A4 sheet.

The next step after obtaining **single kernel crops** is to apply a **background subtraction** technique to extract the area corresponding to the grain in the image. The background subtraction is performed in the HSV color space to obtain a result robust to illumination changes—background threshold has been adjusted for each variety of grains considered in this work according to the background and illumination conditions. Finally, morphological operators are applied to refine and improve contours by eliminating shadows in the scene. The resulting mask is then used to extract the points that define the contour of a given grain. Figure 3$(middle)$ shows illustrations of the cropped regions as well as the results after background subtraction.

The generation of **synthesized images of a cluster of grains** is finally performed by distributing the number of single grains in an empty background image with a homogeneous color. This generation process allows to set different parameters according to the requirements of the final user: the number of grains, size of the resulting synthesized image, background color, percentage of grains in contact, percentage of grains on the image borders (i.e., cut grains) and the number of images to be generated. The algorithm receives as an input a set of images of individual grains randomly selected. Then, based on the number of grains previously defined by the user, and according to a scale factor that takes into account the size of the synthesized cluster image, kernels are randomly rotated and placed one by one in available spaces, ensuring that the grains do not overlap each other. This algorithm does not maximize the contact perimeter between elements, it only ensures that there is a contact according to the random rotation. Since the kernel selection is randomly performed, a grain sample could be considered twice, but by sure with a different orientation in the final image. Figure 3 shows an illustration of the whole pipeline.

4 Experimental Results

This section presents experimental results obtained by using the proposed synthesized image strategy to train instance segmentation approaches. First, the results of a case study that evaluates the performance of the proposed strategy are presented. This case study consists of a set of 23 manually annotated images of corn kernels, the obtained results are considered as a benchmark to compare them with the results obtained when using the synthesized images for training. Then, other categories of grains are included in the evaluation. In this second case, since there are no annotated instances (object's contour), the number of detected instances is considered as the evaluation criteria.

The Mask R-CNN network [7] was trained to generate a model that allows obtaining all the instances of grains present in a given image. The Mask R-CNN network implementation used in this work is based on ResNet-101 as the backbone and pretrained COCO weight. In addition, the images in the training dataset have been resized to 1024×1024, to reduce the computational cost of the entire process. The number of images used in the different datasets has the following distribution: 16 images for training, 4 images for validation, and 3 images for testing. Figure 1 shows the architecture of the Mask R-CNN network used in this work.

The results obtained by training the Mask R-CNN network using different datasets are evaluated in the Sect. 4.2. The metrics used to measure the performance of the trained networks are IoU, the number of grain instances correctly detected, average precision (AP) in IoU 50% (AP_{50}), 75% (AP_{75}), and the average value of IoU 50% to 95% with a step size of 5% (AP@[.5:.95]). Furthermore, the synthesized datasets generated for the different grain varieties have been considered in the training process by using two approaches: (i) first, the Mask R-CNN is individually trained for each grain variety—i.e., single-grain approach; and (ii) the Mask R-CNN is trained considering all the grain varieties at once—i.e., multi-category grain approach.

4.1 Case Study

The results obtained by training the Mask R-CNN network using a real and synthesized clusters of corn kernels datasets are evaluated in this section. The performance of the different schemes (single- and multi-grains) is evaluated as follows: i) by taking into account the number of grain instances correctly counted; ii) by means of the IoU; and iii) through the average precision metric. Table 1 shows experimental results

Table 1. Results on *testing images* (manually annotated ground truth) when the Mask R-CNN network is trained with: Real images (Re); Synthesized single category grain dataset (Sn); Synthesized multi-category grain dataset (Ml)—GT: Ground Truth.

Testing	# of instances				IoU		
images	GT	Re	Sn	Ml	Re	Sn	Ml
Image 1	200	199	198	198	0.901	0.914	0.905
Image 2	190	189	188	187	0.897	0.911	0.902
Image 3	223	215	215	212	0.898	0.900	0.895
Avg	613	**603**	601	593	0.899	**0.908**	0.901

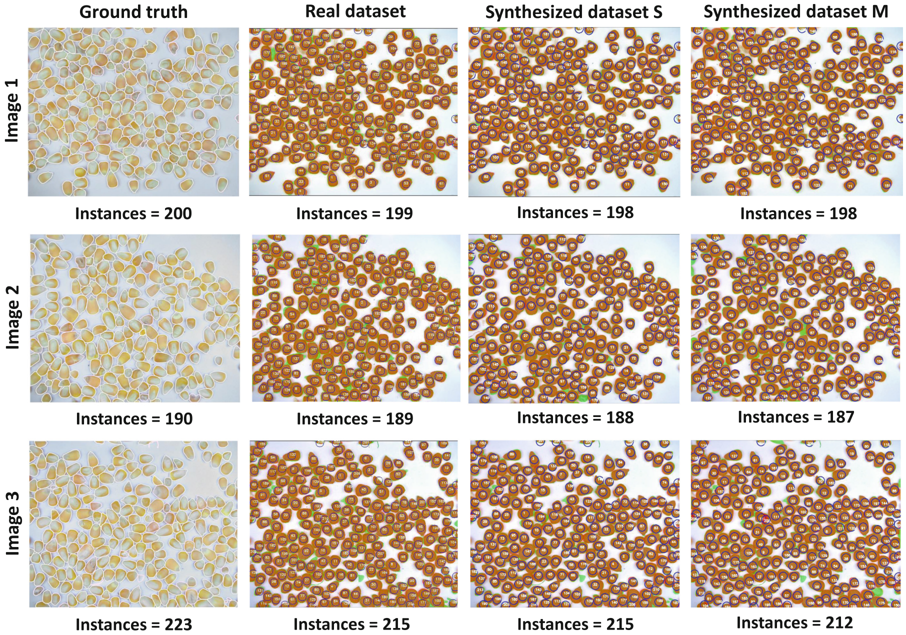

Fig. 4. Results obtained on *testing images* (manually annotated ground truth images) when Mask R-CNN is trained with real and synthesized datasets. (*1st col*) Ground truth labeled with Labelbox. (*2nd col*) Results obtained when training with the real image dataset. (*3rd col*) Results obtained when training is performed with the synthesized single category grain dataset. (*4th col*) Results obtained when training is performed with the synthesized multi-category grain dataset.

obtained with the proposed strategy. The obtained number of instances and the IoU metric computed on a set of testing image datasets (three manually annotated images considered as ground truth) are presented; GT column corresponds to the ground truth number of instances per image; Re columns show the number of predicted instances and IoU metric when the Mask R-CNN network is trained with the real image dataset (manually-annotated images). Sn columns show the number of instances predicted by the network, as well as the IoU metric when trained with the synthesized corn kernels dataset; just the corn category is considered. Finally, the Ml columns correspond to the results obtained when the Mask R-CNN is trained with all synthesized image datasets, in other words, when all grain categories are considered; both numbers of instances and IoU are depicted. Looking at the results depicted in the table, although the number of instances is slightly better when real images are used for training (just one instance error in the first and second testing images) it does not happen the same in the case of IoU metric. The results of the IoU metric show a better performance, in all the cases, when the synthesized dataset is considered (single category grain), an improvement of up to almost 1.4% can be observed in the first testing image. On the other hand, AP results for this case study are depicted in Table 2, where the AP@[.5:.95], AP_{75} and AP_{50} metric values are shown. It can be seen that the Mask R-CNN trained with synthesized images presents a better performance in the metrics AP@[.5:.95] and AP_{75} while the real dataset presents a better result in the metric AP_{50}. The results obtained in the IoU and AP metrics show that the use of synthesized datasets allows with a high percentage of accuracy to correctly delimit the area and contour of the corn kernels confirming the effectiveness and validity of the proposed approach.

Finally, qualitative results and ground truth annotations on the three testing images are shown in Fig. 4, where the number of instances predicted by Mask R-CNN and ground truth values is also depicted. In order to facilitate the qualitative evaluation, the area of each grain segmented by the Mask R-CNN is brown colored while manual annotations are shown in green. In addition, a blue circle has been used to highlight each individual instance together with the corresponding instance number, to check there are no duplicate or bad segmented grains.

Table 2. Results using the AP metric on *testing images* (manually annotated ground truth) when the Mask R-CNN network is trained with: Real images (Re); Synthesized single category grain dataset (Sn); Synthesized multi-category grain dataset (Ml).

Testing	AP@[.5:.95]			AP_{75}			AP_{50}		
images	Re	Sn	Ml	Re	Sn	Ml	Re	Sn	Ml
Image 1	0.790	0.830	0.790	0.964	0.980	0.957	0.989	0.980	0.974
Image 2	0.800	0.830	0.800	0.978	0.984	0.926	0.995	0.989	0.945
Image 3	0.780	0.790	0.750	0.950	0.945	0.884	0.964	0.964	0.903
Avg	0.793	**0.818**	0.781	0.958	**0.964**	0.922	**0.982**	0.978	0.9410

4.2 Free Annotation Results

In order to evaluate the usefulness of the proposed approach in other grain categories, the Mask R-CNN network has been trained with synthesized images generated with the grain types presented in Sect. 3.2. In all these cases, the performance of the network trained with different schemes (single- and multi-grains) is evaluated using the IoU and AP metrics together with the number of correctly detected instances. It should be mentioned that in all these categories of grains (except corn) there are not manually annotated ground truths, hence in the case of real images just qualitative illustrations are depicted together with the number of detected instances, which is used as a quantitative evaluation. Table 3 shows results (i.e., number of instances, IoU, and the AP metrics) obtained when the Mask R-CNN network is trained with the synthesized single category grain dataset (Sn) and with the synthesized multi-category grain dataset (Ml). The number of instances in the GT column corresponds to the total number of grains of the whole testing image sets, while IoU and AP metrics are average values for the whole testing image sets. Three testing images have been used per category; these images contain a random number of instances. It can be appreciated that in most of the cases the best results are obtained when the Mask R-CNN is trained with the synthesized single category grain dataset. Finally, in order to evaluate the performance on real images, Table 4 shows results on different grain categories, just the number of instances, when the Mask R-CNN is trained with synthesized single- and multi- categories schemes, is depicted since there are not manual annotations. Just as illustrations of the performance on real images, Fig. 5 shows segmentation results on real images, obtained by the Mask R-CNN trained with synthesized single category grain dataset for different types of grains.

Table 3. Evaluation results—IoU and AP metrics—of the Mask R-CNN network when trained with synthesized single- and multi- categories; just three testing images per category of synthesized grain clusters are considered. Testing images contain a random number of instances, the total number of instances per category, adding up the three testing images, is depicted in the second column. Ground truth (GT); Synthesized single category grain dataset (Sn); Synthesized multi-category grain dataset (Ml).

Type of grain	# of instances			IoU		AP@[.5:.95]		AP$_{75}$		AP$_{\mathbf{50}}$	
	GT	Sn	Ml	Sn	Ml	Sn	Ml	Sn	Ml	Sn	Ml
Bean	973	973	973	0.943	0.937	0.918	0.884	0.999	0.998	0.999	0.999
Chickpea	903	903	903	0.936	0.928	0.928	0.643	0.999	0.998	0.999	0.981
Corn	598	598	598	0.942	0.934	0.929	0.886	0.998	0.997	0.997	0.988
Lentil	913	912	912	0.946	0.936	0.917	0.907	0.999	0.998	0.999	0.999
Mote	840	840	840	0.935	0.929	0.917	0.904	0.999	0.998	0.999	0.999
Soybean	1836	1835	1835	0.906	0.902	0.836	0.830	0.999	0.998	0.999	0.999
Sunflower	768	768	768	0.928	0.923	0.899	0.876	0.999	0.998	0.999	0.998

Table 4. Evaluation results on real-world grain cluster testing images: number of instances obtained by the Mask R-CNN network when trained with synthesized single- and multi- categories. Ground truth (GT); Synthesized single category grain dataset (Sn); Synthesized multi-category grain dataset (Ml).

Type of	Testing	# of instances		
grain	images	GT	Sn	Ml
Bean	13	1890	1888	1887
Chickpea	23	2871	2864	2868
Corn	9	1388	1387	1387
Lentil	9	2142	2137	2140
Mote	23	2985	2973	2969
Soybean	11	2500	2494	2484
Sunflower	22	3000	2829	2751

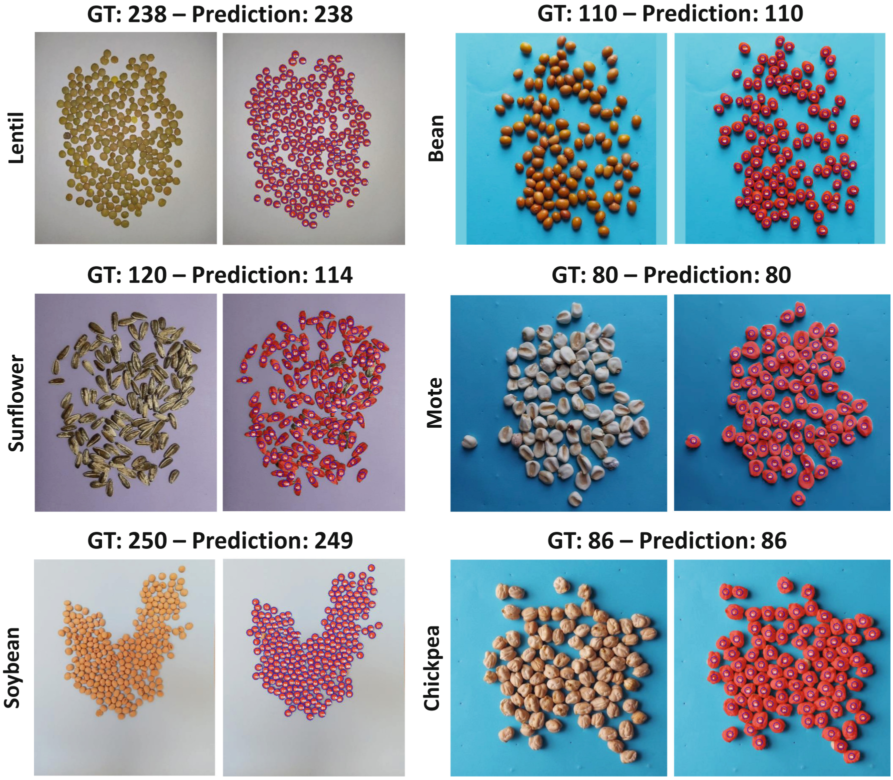
GT: 238 – Prediction: 238
Lentil
GT: 110 – Prediction: 110
Bean
GT: 120 – Prediction: 114
Sunflower
GT: 80 – Prediction: 80
Mote
GT: 250 – Prediction: 249
Soybean
GT: 86 – Prediction: 86
Chickpea

Fig. 5. Illustrations (one case per grain category) on real images from the Mask R-CNN trained with the corresponding synthesized single category grain dataset.

5 Conclusions

This paper proposes a simple but efficient strategy to automatically obtain annotations to be used in the grain segmentation problem. Although it has been evaluated with the Mask R-CNN architecture, it can be used in other deep learning based approaches. The corn kernel case study shows that the Mask R-CNN network trained with the proposed synthesized datasets achieves similar or better results, both IoU and AP, than when trained with manually annotated images. The simplicity of the proposed strategy allows generating ground truth information (annotated set of instances) just by taking a set of images with instances regularly distributed in a grid; in other words, the time-consuming annotation task is not required, speeding up the training process and at the same time reaching better results. It should be mentioned that the results obtained by the proposed strategy can be easily improved by just increasing the number of instances initially acquired (regular grid), or increasing the variability of considered grains. In other words, there is still space for improvement. In addition to the case study, the proposed strategy has been evaluated with other types of grains, which include different shapes, textures, and background colors (e.g., lentil, sunflower, bean, mote, chickpea, and soybean). In all the cases the proposed strategy shows its validity. Again, results can be improved by enlarging the synthesized images in the dataset used for training as well as the variability of instances. Finally, just to confirm the need of having annotation for each type of grains, a dataset with annotations of all the classes has been evaluated, in all the cases the single class case reaches better results—i.e., the Mask R-CNN trained with the grain type to be considered.

Acknowledgements. This work has been partially supported by the ESPOL Polytechnic University; the Spanish Government under Project TIN2017-89723-P; and the "CERCA Programme/Generalitat de Catalunya". The authors gratefully acknowledge the support of the CYTED Network: "Ibero-American Thematic Network on ICT Applications for Smart Cities" (REF-518RT0559) and the NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

1. Bai, M., Urtasun, R.: Deep watershed transform for instance segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5221–5229 (2017)
2. Bolya, D., Zhou, C., Xiao, F., Lee, Y.J.: YOLACT: real-time instance segmentation. In: IEEE/CVF International Conference on Computer Vision, pp. 9157–9166 (2019)
3. Chabrier, S., Emile, B., Rosenberger, C., Laurent, H.: Unsupervised performance evaluation of image segmentation. J. Adv. Signal Process. **2006**, 1–12 (2006). https://doi.org/10.1155/ASP/2006/96306
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)
5. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010). https://doi.org/10.1007/s11263-009-0275-4

6. Fedorov, A., Nikolskaia, K., Ivanov, S., Shepelev, V., Minbaleev, A.: Traffic flow estimation with data from a video surveillance camera. J. Big Data **6**(1), 1–15 (2019). https://doi.org/10.1186/s40537-019-0234-z
7. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
9. Kar, A., Kulshreshtha, P., Agrawal, A., Palakkal, S., Boregowda, L.R.: Annotation-free quality estimation of food grains using deep neural network. In: 30th British Machine Vision Conference, pp. 1–12 (2019)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
11. Levesque, V.: Texture segmentation using Gabor filters, pp. 1–8. Center for Intelligent Machines. McGill University (2000)
12. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part V. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
13. Mao, J., Xiao, T., Jiang, Y., Cao, Z.: What can help pedestrian detection? In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3127–3136 (2017)
14. Poma, X.S., Riba, E., Sappa, A.: Dense extreme inception network: towards a robust CNN model for edge detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1923–1932 (2020)
15. Radovic, M., Adarkwa, O., Wang, Q.: Object recognition in aerial images using convolutional neural networks. J. Imaging **3**(2), 1–21 (2017)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations, pp. 1–14 (2015)
17. Toda, Y., Okura, F., Ito, J., Okada, S., Kinoshita, T., Tsuji, H., Saisho, D.: Training instance segmentation neural network with synthetic datasets for crop seed phenotyping. Commun. Biol. **3**(1), 1–12 (2020)
18. Velesaca, H.O., Mira, R., Suarez, P.L., Larrea, C.X., Sappa, A.D.: Deep learning based corn kernel classification. In: IEEE International Conference on Computer Vision and Pattern Recognition Workshops, pp. 294–302 (2020)
19. Wang, Y., Sun, S.: Image-based grain partitioning using skeleton extension erosion method. J. Pet. Sci. Eng. **205**, 1–11 (2021)
20. Wei, J., He, J., Zhou, Y., Chen, K., Tang, Z., Xiong, Z.: Enhanced object detection with deep convolutional neural networks for advanced driving assistance. IEEE Trans. Intell. Transp. Syst. **21**(4), 1572–1583 (2019)
21. Wofk, D., Ma, F., Yang, T.J., Karaman, S., Sze, V.: FastDepth: fast monocular depth estimation on embedded systems. In: IEEE International Conference on Robotics and Automation, pp. 6101–6108 (2019)
22. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: large-scale scene recognition from abbey to zoo. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3485–3492 (2010)
23. Yang, S., Wang, W., Liu, C., Deng, W.: Scene understanding in deep learning-based end-to-end controllers for autonomous vehicles. IEEE Trans. Syst. Man Cybern. Syst. **49**(1), 53–63 (2018)
24. Yu, Y., Zhang, K., Yang, L., Zhang, D.: Fruit detection for strawberry harvesting robot in non-structural environment based on Mask R-CNN. Comput. Electron. Agric. **163**, 1–9 (2019)