Centroid Loss for Weakly-Supervised Semantic Segmentation in a Quality-Control Application

Kai Yao, Alberto Ortiz, Francisco Bonnin-Pascual
Department of Mathematics and Computer Science, *University of the Balearic Islands* and IDISBA, Palma, Spain
yaokaimallorca@gmail.com, {alberto.ortiz, xisco.bonnin}@uib.es

***Abstract*—Process automation is enabling a level of accuracy and productivity that goes beyond human ability, and one critical area where automation is making a big difference is quality control. In this paper, we describe a semantic segmentation solution aiming at detecting the presence of quality control elements in surgery toolboxes prepared by the sterilization unit of a hospital. In order to reduce the time required to prepare pixel-level ground truth, this work focuses on the use of weakly-supervised annotations (scribbles). Moreover, our solution integrates a clustering approach into a semantic segmentation network, thereby reducing the negative effects caused by weakly-supervised annotations. The paper describes the design process and reports on the results obtained.**

***Index Terms*—Quality Control, Object Recognition, Weakly-Supervised Semantic Segmentation**

I. Introduction

Process automation is enabling a level of accuracy and productivity that goes beyond human ability, and one critical area where automation is making a big difference is quality control. Machine vision and deep learning are nowadays changing the game for this kind of application. Generically speaking, these technologies enable inspection automation for every product on the line, and this means consistent and accurate results, and, correspondingly, direct contribution to profits from the reduction of the percentage of failing operations.

Furthermore, enabling non-contact, thus non-destructive inspection, optical techniques are especially well suited when the correct manipulation of the object under inspection is crucial. This is precisely the inspection problem that we deal with in this paper: it consists in the detection of a number of control elements that the sterilization unit of a hospital places in boxes and bags containing surgical tools that surgeons and nurses have to be supplied with prior to starting surgery. These elements provide evidence that the tools have been properly submitted to the required cleaning processes. Figure 1 shows, from left to right and top to bottom, the six kinds of elements to be detected for this application: the label/bar code used to track a box/bag of tools, the yellowish seal, the three kinds of paper tape which changes to the black-, blue- and pink-stripped appearance when the box/bag has been inside the autoclave, and an internal filter which is placed inside some boxes and creates the white-dotted texture that can be observed (instead of black-dotted when the filter is not inside).

Figure 1. Objects to be detected: label, seal, 3 kinds of paper tape (black-, blue- and pink-stripped) and the internal filter. As can be observed in the top-left case, one can find several of these items in the same image.

In a previous work [1], we developed an arbitrarily-oriented object detector based on DCNNs for oriented bounding boxes regression. Despite the general good performance exhibited by this detector (as already reported), when lots of objects appear in a small area, bounding boxes-based object recognition tends to degrade in performance, and also, most importantly, detection results become less informative and even messy. Due to this reason, we have considered the adoption of an alternative semantic segmentation approach to detect targets at the pixel level, increasing thus the localization accuracy. This in particular means the availability of pixel-level manual annotation of objects for the training set, what is a very high time-consuming task, especially when the dataset comprises a large number of images. In this regard, in this work, we address the development of the detector from a weakly-supervised learning perspective, using simple scribbles to simplify the preparation of the training set (useful during exploitation stage when updates have to be performed). An example on the use of scribbles as ground truth data is shown in Fig. 2 [left].

Unlike previous works such as [2]–[4], that perform weakly-supervised image segmentation through a multi-level network, our method employs an end-to-end architecture and a multi-task joint training strategy. The main contributions of this work are described in Section II, namely: (1) we propose a centroid loss function for weakly-supervised semantic segmentation and evaluate its performance for our task; (2) we design a Mean Square Error (MSE)-based regularization term based on the predicted centroids in order to improve the segmentation

This work is partially supported by EU-H2020 projects BUGWRIGHT2 (GA 871260) and ROBINS (GA 779776), project PGC2018-095709-B-C21 (MCIU/AEI/FEDER, UE), and by project PROCOE/4/2017 (Govern Balear, 50% P.O. FEDER 2014-2020 Illes Balears). This publication reflects only the authors views and the European Union is not liable for any use that may be made of the information contained therein.

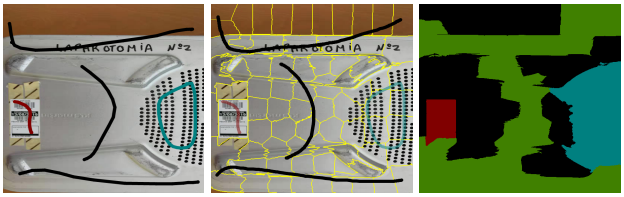


Figure 2. Illustration of the pseudo-masks generation process from scribbles-based ground truth and superpixels: [left] scribbles for the *label* and *internal filter* classes, [middle] result of the SLICO superpixels algorithm, and [right] resulting pseudo-mask.

performance. To finish, the experimental results reported in Section III show that our approach achieves good performance for the intended task.

II. Methodology

In this section, we will discuss the methodology underlying our approach. For a start, in Section II-A, we refer to the generation of pseudo-masks for training on the basis of the output of a superpixel-segmentation algorithm and scribbles. The architecture of the DCNN is described next in Section II-B. Sections II-C and II-D discuss on the use of, respectively, the partial cross entropy loss and the centroid loss functions. To finish, the final loss function is presented in Section II-E.

A. Pseudo-mask Generation by means of Superpixels

Figure 2[left] illustrates the use of scribbles for image annotation, where the red scribble denotes the *label* class, the blue scribble is for the *internal filter* class and the black scribbles correspond to background. It is clear that scribbles already label some pixels of each class, which could be used for training. However, unfortunately, the performance of the network in this case turns out to be very poor. Inspired by ScribbleSup [5], to address the latter problem we generate training pseudo-masks using the superpixels produced by Adaptive-SLIC (SLICO) [6] (see Fig. 2[middle]), so that, taking advantage of the good boundary adherence of superpixels, training pixels belonging to a superpixel intersecting with a scribble are labelled with the same class as the scribble (see Fig. 2[right], where green pixels denote the background, red pixels are for the *label* class, blue pixels are for the *internal filter* class and the black pixels denote unlabelled data, which are ignored during training).

B. Network Architecture

In this work, similarly to [7], we embed attention modules in U-Net, a popular encoder-decoder DCNN devised for biomedical images, to improve its ability to segment small targets. In more detail, *attention gates* (AGs) are integrated into the decoding part of U-Net, where, as shown in Fig. 3, an AG is fed by two input tensors, one from the encoder and the other from the decoder. Unlike the *Squeeze-and-Excitation* (SE) block of [8], which obtains attention weights for filter selection, we use AGs to compute attention weights for pixels.

Additionally, we integrate in our Attention U-Net (AUN) a sub-net to compute the centroid loss (see Section II-D). In comparison with AUN, the network of Fig. 3: (1) handles two

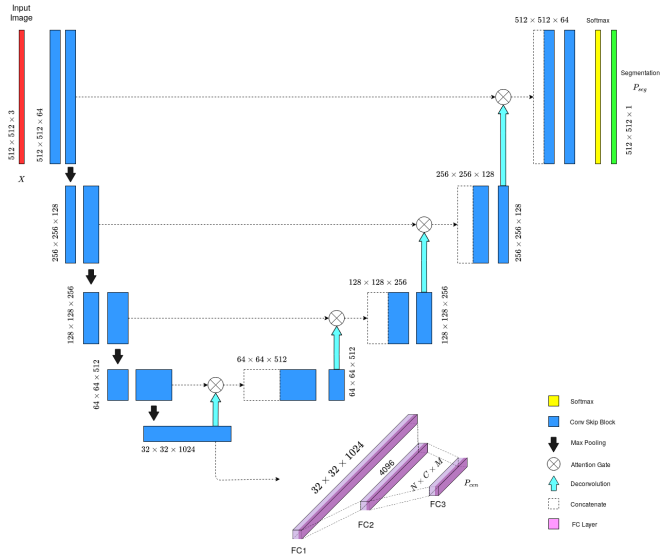


Figure 3. Schemtic description of the Centroids AUN model. Size decreases gradually by a factor of 2 at each scale in the encoding part, and increases by the same factor in the decoding part. In the latter, we use AGs to help the network focus on areas of high-response in the feature maps. The *Conv Skip* block is the *skip connection* of ResNet [9]. In this figure, N is the batch size, C is the number of classes and M is the clustering feature space dimension.

sorts of ground truth during training, scribbles Y_{scr} to train the sub-net for proper centroids prediction and pseudo-masks Y_{seg} for segmentation training; and (2) produces two outputs, centroids P_{cen} and the image segmentation P_{seg}. In order to avoid massive computation and save memory, we embed the sub-net in the intermediate layers of the network, instead of at the end. As shown in Fig. 3, the sub-net comprises three blocks, and every block includes a fully connected layer, a batchnorm layer and a ReLU activation layer.

C. Partial Cross Entropy Loss Function

Given a C-class problem and a training set Ω, comprising a subset Ω_L of labelled pixels and a set Ω_U of unlabeled pixels, the *Partial Cross Entropy Loss* L_{pCE}, widely used in weakly-supervised image segmentation, computes the cross entropy only for labeled pixels $p \in \Omega_L$, ignoring $p \in \Omega_U$:

$$L_{\text{pCE}} = \sum_{c=1}^{C} \sum_{p \in \Omega_L} -y_{g(p),c} \log y_{s(p),c} \,, \tag{1}$$

where $y_{g(p),c} \in \{0, 1\}$ is the ground truth and $y_{s(p),c}$ is the segmentation result. In our case, and for L_{pCE}, we define $\Omega_L^{(1)}$ as the pixels labelled in pseudo-masks, while $y_{s(p),c}$ is as supplied by the softmax final network layer.

D. Centroid Loss Function

Although the performance of the network when trained using pseudo-masks is not bad, we have noticed that segmentation performance depends on the quality of the pseudo-masks and hence on the quality of the superpixels, i.e. how they adhere to class boundaries. The *Centroid Loss* function is introduced in this section precisely to avoid this dependence and improve segmentation results.

To this end, this loss function actually implements a clustering process similar to K-means. Briefly speaking, K-means iteratively calculates a set of centroids μ_c for the considered number of clusters/classes, associating samples to closest clusters in feature space, so as to minimize the intra-class variance until convergence. Contrarily to other CNN-based clustering approaches, which reformulate K-means as a neural network optimizing the intra-class variance loss by means of a backpropagation-style scheme [10], [11], in this work we define the centroid loss L_{cen} as a partial cross-entropy loss considering in this case $\Omega_L^{(2)}$ as the set of pixels coinciding with the scribbles:

$$L_{\text{cen}} = \sum_{c=1}^{C} \sum_{p \in \Omega_L^{(2)}} -y_{g(p),c} \log y'_{s(p),c} \tag{2}$$

$$y'_{s(p),c} = \frac{\exp(d_{p,c})}{\sum_{c'=1}^{C} \exp(d_{p,c'})}, \; d_{p,c} = \frac{||f_p - \mu_c||_2^2}{\sum_{c'=1}^{C} ||f_p - \mu_{c'}||_2^2}$$

where: (1) f_p is the feature of pixel p and is obtained from the last convolutional layer of the encoding part of the network and (2) μ_c is the centroid predicted for class c, i.e. $\mu_c \in P_{\text{cen}}$ (see Fig. 3 for 1 and 2). With equation 2, we transform the process of minimizing the distances from samples to centroids to a process of searching for the clustering with highest probability thanks to the softmax formulation adopted.

E. Full Loss Function

Since L_{pCE} only applies to pixels labelled in the pseudo-mask and L_{cen} is also restricted to a subset of image pixels, namely pixels coinciding with scribbles, we add a third loss term in the form of a normalized MSE loss L_{mse} as a regularization term involving all pixels p for which a class label is predicted $\Omega_L^{(3)}$. Similarly to L_{cen}, we define L_{mse} in terms of softmax-normalized Euclidean distances:

$$L_{\text{mse}} = \frac{\sum_{c=1}^{C} \sum_{p \in \Omega_L^{(3)}} d_{p,c}}{N \cdot C \cdot |\Omega_L^{(3)}|} \tag{3}$$

where N is the batch size and $|\mathcal{A}|$ stands for the cardinality of set $\mathcal{A}$.

In the end, the complete loss function, to be calculated for every image in the training batch, is:

$$L = L_{\text{pCE}} + \lambda_{\text{cen}} L_{\text{cen}} + \lambda_{\text{mse}} L_{\text{mse}} \tag{4}$$

where λ_{cen} and λ_{mse} are trade-off constants.

III. Experimental Results

The following sections describe the experimental setup (Section III-A) and the results obtained through several experiments aiming at exploring the performance of the semantic segmentation approach described in this paper (Section III-B).

Table I
mIOU for different loss functions

	L_{pCE}	L_{cen}	L_{mse}	mIOU
E1	✓			0.6770
E2	✓	✓		0.7594
E3	✓	✓	✓	0.7679

A. Experimental Setup

We have employed a dataset comprising 484 images, which, as usual, has been split in a training set (2/3) and a testing set (1/3). The scribble annotations have been generated by means of 10-pixel brushstrokes, avoiding the annotation of object boundaries or ambiguous regions. To generate the pseudo-masks, we have configured SLICO to produce 100 superpixels/image. Performance results are reported in the next section as the *mean Intersection Over Union* (mIOU), using the fully supervised ground truth. All experiments have been conducted within the Pytorch framework, running Ubuntu 64-bit on a desktop PC fitted with a 2.9GHz 12-core CPU with 32 Gb RAM and an NVIDIA GeForce RTX 2080 Ti GPU. In all the experiments, we set constants λ_{cen} and λ_{mse} to 1. The batch size N is 6 and the input images are resized to 512×512 pixels, which is the best configuration for our GPU.

B. Performance Results

To illustrate the performance of our approach, we run experiments E1-3 described in Table I, where we are adding terms to the loss function until reaching the form of equation 4. In all cases, the dimension of feature space during clustering M is set to 7 (experimentally found).

a) Effect of Centroid Loss: Figure 4[left] plots L_{pCE} and L_{cen} until epoch 100. As can be observed, the network manages to minimize L_{cen} by a large margin in both E2 and E3. Regarding L_{pCE}, curve L_{pCE} for E1 is always above the other two curves for E2-3, what means that the centroid loss L_{cen} contributes positively during training, decreasing L_{pCE}. Additionally, the mIOU curves of E2 and E3 are above the mIOU curve of E1, so that L_{cen} leads to better performance also for the test set. Final mIOU results confirming this trend can be found in Table I, with mIOU 0.7594 and 0.7679 for, respectively, E2 and E3, and 0.6770 for E1.

b) Effect of MSE regularization: Figure 4[right] shows the change of L_{mse} during training. Unlike L_{pCE} and L_{cen} curves for E1-3, which decrease notoriously during training, L_{mse} exhibits a trend from significant fluctuation to gradual stability without decreasing. The reason is that, at the beginning of the training, the network predictions are mostly erroneous, what leads to the curve to fluctuate dramatically. But, as the training progresses, the network predictions improve and the L_{mse} becomes more stable. As a result, E3 obtains the highest mIOU (0.7679), as shown in Table I.

c) Qualitative comparison: To finish, Fig. 5 compares visually the scribbles and pseudo-masks of departure, and the resulting segmentations for two images and cases E1-3. As can be seen, the centroid loss can decrease the negative effect of a noisy labelling in the respective pseudo-masks, outperforming

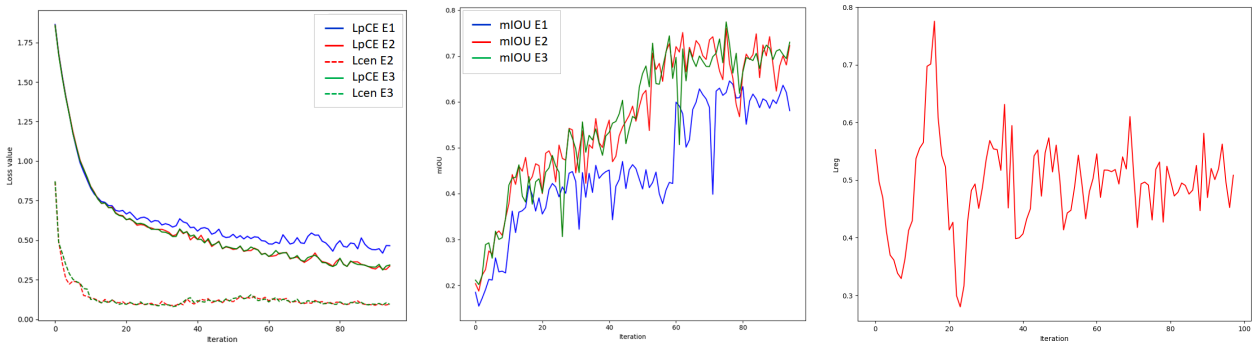
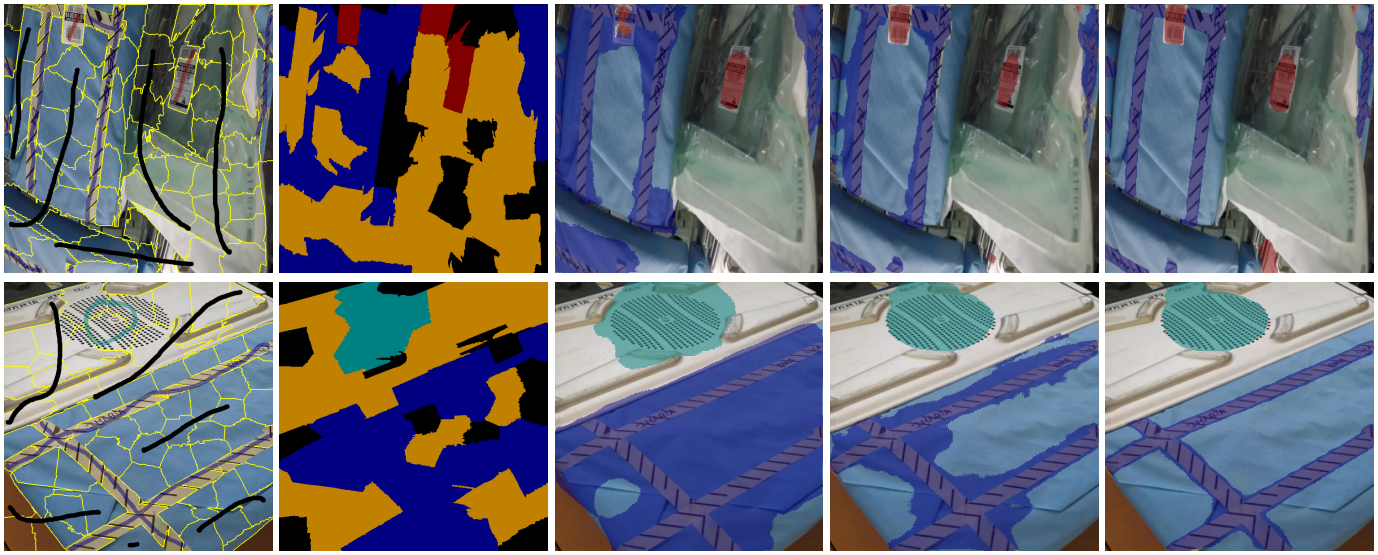


Figure 4. Training curves during E1-3: [right] loss value for L_{pCE} and L_{cen}, [middle] change in mIOU for the test set, and [right] L_{mse} for E3



scribbles & superpixels | pseudo-mask | E1 | E2 | E3

Figure 5. Examples of segmentation results: [1st col.] scribble ground truth and superpixel segmentation; [2nd col.] resulting pseudo-mask, where the black pixels denote unlabeled data, orange pixels denote the *background* class, blue pixels denote the *black-stripped paper tape* class, red pixels are for the *label* class and light blue for the *internal filter* class; [3rd-5th col.] segmentation results for the respective cases.

a network trained only by means of L_{pCE}, as shown for E1 and E2 in Fig. 5. Segmentation results for E3, on the other side, show that the MSE regularization term contributes to a better adherence of the resulting segmentations to class boundaries.

IV. Conclusion and Future Work

We have proposed a weakly-supervised segmentation approach based on AUN for an object recognition application. The loss function comprises three terms based on convenient scribble annotations, which are jointly optimized within an end-to-end model. As has been reported in the experimental results section, our approach achieves a competitive performance at a low cost as for ground truth labelling. As for future work, we plan to follow the same research line and develop other types of weak annotations and training strategies, so as to achieve higher performance and simpler annotations.

References

[1] K. Yao, A. Ortiz, and F. Bonnin-Pascual, "A DCNN-based Arbitrarily-Oriented Object Detector for a Quality-Control Application," in *ETFA*. IEEE, 2019, pp. 1507–1510.

[2] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, and J. Jiao, "Weakly supervised instance segmentation using class peak response," in *CVPR*, 2018, pp. 3791–3800.

[3] X. Wang, S. You, X. Li, and H. Ma, "Weakly-supervised semantic segmentation by iteratively mining common object features," in *CVPR*, 2018, pp. 1354–1362.

[4] L. Chan, M. Hosseini, and K. Plataniotis, "A comprehensive analysis of weakly-supervised semantic segmentation in different image domains," *arXiv*, 2020. [Online]. Available: http://arxiv.org/abs/1912.11186

[5] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *CVPR*, 2016, pp. 3159–3167.

[6] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC Superpixels Compared to State-of-the-Art Superpixel Methods," *PAMI*, vol. 34, no. 11, pp. 2274–2282, Nov 2012.

[7] O. Oktay *et al.*, "Attention U-net: Learning where to look for the pancreas," *arXiv*, 2018. [Online]. Available: http://arxiv.org/abs/1804.03999

[8] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018, pp. 7132–7141.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[10] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*, 2016, pp. 499–515.

[11] X. Peng, J. T. Zhou, and H. Zhu, "k-meansNet: When k-means meets differentiable programming," *arXiv*, 2019. [Online]. Available: http://arxiv.org/abs/1808.07292