Learning to Segment Fine Structures Under Image-Level Supervision With an Application to Nematode Segmentation*

Long Chen[1], Martin Strauch[1], Matthias Daub[2], Hans-Georg Luigs[3], Marcus Jansen[3] and Dorit Merhof[1]

***Abstract*— Image segmentation models trained only with image-level labels have become increasingly popular as they require significantly less annotation effort than models trained with scribble, bounding box or pixel-wise annotations. While methods utilizing image-level labels achieve promising performance for the segmentation of larger-scale objects, they perform less well for the fine structures frequently encountered in biological images. In order to address this performance gap, we propose a deep network architecture based on two key principles, Global Weighted Pooling (GWP) and segmentation refinement by low-level image cues, that, together, make segmentation of fine structures possible. We apply our segmentation method to image datasets containing such fine structures, nematodes (worms + eggs) and nematode cysts immersed in organic debris objects, which is an application scenario encountered in automated soil sample screening. Supervised only with image-level labels, our approach achieves Dice coefficients of 79.72% and 58.51 % for nematode and nematode cyst segmentation, respectively.**

I. Introduction

Deep learning models have recently brought tremendous progress in automatic image segmentation tasks. However, collecting pixel-level annotations, which is required in the fully-supervised setting, is time-consuming and labor-intensive. The problem is particularly prominent in the biomedical field due to the expertise required for annotation. To alleviate the demand for pixel-level annotations, many weakly-supervised approaches were proposed [1], [2], [3] to train segmentation models with dot annotations, scribbles or bounding boxes. Among the variants of weak supervision, image-level supervision is particularly appealing due to the efficiency of obtaining training data. Although quite promising performance has been achieved on natural image and urban scene datasets [4], [6], [5], such as Pascal VOC and Cityscapes, image-level supervision is rarely studied for the segmentation of biomedical images that often contain fine-scale structures.

State-of-the-art methods [4], [5], [6] follow two main principles: (1) estimate a coarse segmentation from the class-specific activation map [7], [8], [9] and (2) refine the coarse segmentation using low-level image processing techniques, such as denseCRF [10]. Recent research [7], [8], [9] has shown that convolutional layers spontaneously localize objects of interest despite only being trained through image-level classification. In the classification activation map (CAM) approach [7], convolutional feature maps are spatially pooled with Global Average Pooling (GAP) before the last fully connected layer. The class-specific activation map is essentially the sum of convolutional feature maps weighted by parameters of a certain class in the fully connected layer. Grad-CAM [8] generalizes the CAM [7] approach by weighting convolutional feature maps with gradients, so that the network structure is not necessarily limited to the one used by [7].

Here, we introduce a new pooling scheme, Global Weighted Pooling (GWP), a weighted generalization of GAP [7]. In our experiments (Section III), GWP is proved to be clearly better for the segmentation of fine structures in biological images than both Global Max Pooling (GMP) and GAP [7], that fails if the object is very small compared to the image size.

Our application scenario for image segmentation is the automated large-scale screening of soil samples infested with sugar beet nematodes (*Heterodera schachtii*). Nematode are plant parasites that are responsible for considerable financial losses in the agricultural industry [11]. An automated system (name omitted for anonymous review) allows users to record microscopic images of processed soil samples containing nematodes (worms and eggs) and nematode cysts amidst organic debris objects (Figure 1). Here, detection and segmentation is the basis for automated data analysis, including counting and phenotyping the nematodes. We evaluate how different segmentation variants with image-level supervision perform in this application scenario with fine-scale biological structures (Section III).

Main contributions

1) We propose a CNN architecture and training mode for foreground segmentation under *image-level supervision*.
2) Since neither GAP nor GMP are suitable for small objects and fine structures, we propose *Global Weighted Pooling (GWP)* as an alternative.
3) We use an efficient superpixel [12] voting approach to remove noise and to improve the prediction at boundaries. The segmentation is refined by iteratively tuning the network with masks generated by superpixel voting.
4) As important plant parasites, nematodes are of great significance in agricultural monitoring [13]. By eval-

*This work was funded by the Germany Ministry of Education and Research, project KMU-innovativ-19: PheNeSens, grant number 031B0474C

[1]Imaging and Computer Vision, RWTH Aachen University, Aachen, Germany {long.chen, martin.strauch, dorit.merhof}@rwth-aachen.de

[2]Julius Kühn Institute: Federal Research Centre for Cultivated Plants, Elsdorf, Germany matthias.daub@julius-kuehn.de

[2]LemnaTec GmbH, Aachen, Germany {marcus.jansen, hans-georg.luigs}@lemnatec.de

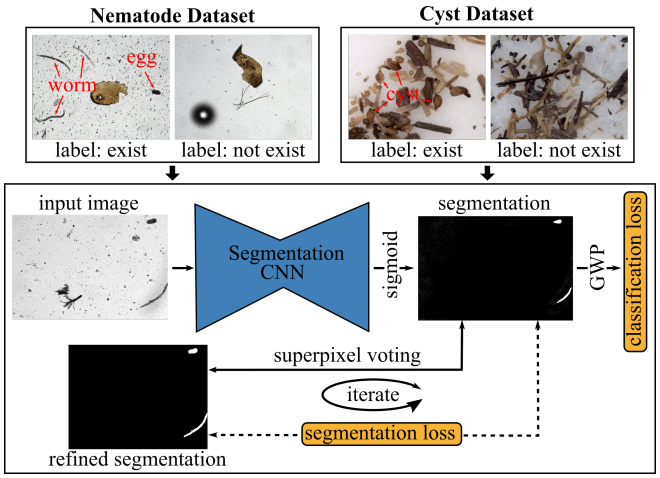

Fig. 1. **Overview of our image-label supervised segmentation model.** The model is firstly trained by minimizing the *classification loss*. With the proposed global weighted pooling (GWP), the model can identify the coarse extent of the object. Further refinement is conducted by alternating mask refinement through superpixel voting and optimizing *segmentation loss* with refined masks as ground truth. The only required manual annotation is the image-level label.

uation on nematode datasets, our work directly contributes to nematode screening and phenotyping applications.

II. Method

Our model is trained in stages with the *classification loss* and *segmentation loss*, respectively (Figure 1). Under the classification supervision, the model learns to localize discriminative regions in the image, that is, segmentation of objects (Section II-A). It is worth mentioning that a proper pooling strategy is critical to successfully identify the extent of small objects.

The initial segmentation is usually very coarse and could be further refined by adjusting segmentation boundaries to image edges. To this end, we propose a pixel voting approach (Section II-B). The model is then iteratively tuned with refined mask as the segmentation ground truth (Section II-C).

A. Pooling class scores form segmentation map

Different from [7], we use a more straightforward setting: one sigmoid activated map for each foreground class. Whether the task is a binary segmentation problem or deals with multiple foreground classes, for each category only two cases need to be considered: Objects belonging to this category present or not. From negative examples, the model learns structures which are not the object of interest, and neurons at all such locations are trained to be silent. Then the activated neurons will indicate the object location, when positive examples are present and stated as positive explicitly.

In case of small objects and fine structures, the pooling strategy is critical to a successful training. Widely used on natural image datasets, the global average pooling (GAP) is reported to be able to estimate the object extent well. However, GAP fails the whole model, when the target object

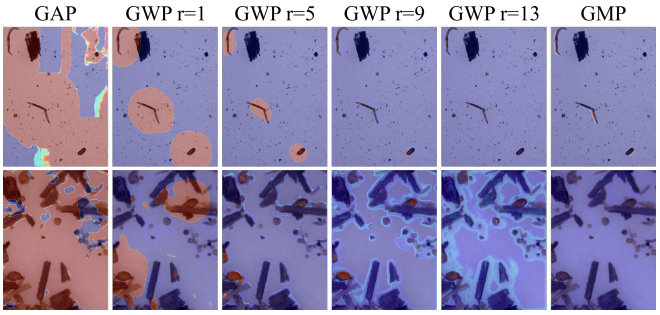

Fig. 2. **Activation map trained with different pooling methods.** The global max pooling (GMP) only activates a few pixels in the boundary area of objects, while the global average pooling (GAP) aggressively highlights a large region. Our proposed global weighted pooling (GWP) can well adapt to the object extent in a large value range of r . The first and second row show the results in images of nematode and cyst.

is very small compared to the image size. This is because there must be a fair amount of neurons activated to allow the average value to reach the activation level, while objects are small.

The global max pooling (GMP) seems to be a feasible option: when a negative image is shown, the spatially maximal value is trained toward zero, therefore all neurons in the segmentaion map are suppressed to be silent. However, GMP tends to only activate a few locations in the positive image, since only one single maximal value contributes to the classification score. In biomedical images, it often happens that only a small part of the pixels on the edge of objects are activated, as shown in Figure 2. A small number of activated pixels in the boundary area may be sufficient for image classification, but not for further segmentation refinement. GMP may result in an activation map that highlights the internal area of objects. This situation did occur very rarely in our experiments. We argue this is related to network initialization. If it happens that the inner area has a higher initial value, the GMP will enhance and highlight this area. Since the inner area, the boundary area and even surrounding area could be discriminative features for image classification, and given the fact that GMP only considers the maximum value, there is no guarantee that an activation map covering a considerable area of the object will be obtained. Therefore, we propose to use the global weighted pooling (GWP):

$$s_{pool} = \frac{\sum_{i=1}^{N}(\exp(rs_i) - 1 + \epsilon)s_i}{\sum_{i=1}^{N}(\exp(rs_i) - 1 + \epsilon)}$$

where s_i is the sigmoid activated value at each location of the segmentation map. The hyper-parameter r controls how the pooling behaves: high r values implies an effect similar to GMP. When $r = 0$, all weights are equal and GWP is exactly the same as GAP. The parameter ϵ is empirically set according to the magnitude order of the reciprocal of image size ($\epsilon = 1e^{-5}$ in this work). The pooled class scores are trained with *crossentropy loss*. See Figure 2 for comparison of activation maps trained with different pooling methods.

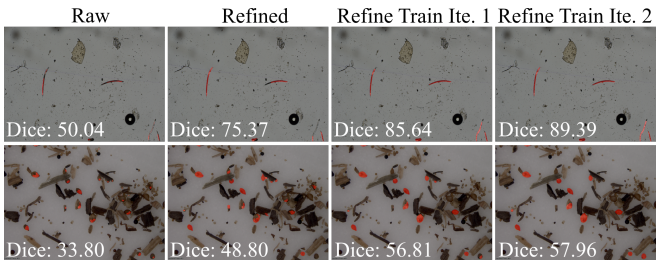

Fig. 3. **Segmentation results in different refinement stages.** The first and second row show the segmentation of nematodes (worm + egg) and nematode cysts, respectively. The model is trained using GWP with $r = 9$ in both cases.

B. Superpixel Voting

Generating either an oversegmentation or undersegmentation (Figure 2), the raw activation map barely aligns with the object boundary spontaneously. Therefore, refinement can be conducted by adjusting the segmentation boundary to close low-level image edges. To this end, we propose a method called superpixel voting. Superpixels are grouping of pixels with similar low-level features and spatially close, therefore, can be regarded as a consistent area either belonging to the foreground or the background.

The pixel voting method first compute all superpixels in an images and then count the proportion of foreground pixels for each superpixel. If the the proportion is above a certain threshold, the whole superpixel is taken as foreground. In this work, we use Felzenszwalb's efficient graph based method [12] to compute superpixels. We set the single parameter scale of Felzenszwalb's method to 100 and use 0.2 voting rate for all experiment, unless mentioned otherwise.

C. Iterative Segmentation Refinement

The segmentation refined by superpixel voting (Section II-B) can further be used as ground truth to finetune the network output. In this way, the superpixel voting is not required any more as a postprocessing step at inference time. In addition, we demonstrate that the tuned network output can outperform its supervision ground truth, the mask refined through superpixel voting. Therefore, the estimated ground truth can be iteratively refined by using the network outputs, while the network can be iteratively finetuned as well. In this work, we use *Dice loss* [14] as the segmentation loss.

III. Datasets and Experiments

A. Datasets

Two datasets used for evaluation was collected from the PheNeSens project, aiming at conducting nematode screening and phenotyping in a high-throughput manner. The cyst sample is obtained from soil samples through physical separation processing, such as washing, filtering and centrifuging. By grinding and suspending the cysts, the content can be extracted for further observation, which is the nematode sample. See Figure 3 for demo images.

Nematode Dataset: We collected 3967 images of 3008x4112 pixels, containing ∼4000 nematode worms and ∼5500 eggs. In addition to nematodes (eggs + worms), some disturbing structures also appear in the image, such as nematode cyst skin, organic debris and liquid bubbles. All nematodes are manually outlined with the help of the intelligent scissors approach [15], which took about 200 hours of work.

Cyst Dataset: The cyst dataset consists of 487 image of size 1000x1240 pixels, with 8855 cysts manually annotated. Compared to the nematode dataset, the cyst dataset is more challenging from the perspective of image processing, with a highly cluttered object collection. The disturbing objects are mainly organic debris of similar size and density to the cysts, such as root fragments and weed seeds, which is barely separable by physical separation processing.

B. Network training

To train the model, positive and negative image examples are required. We randomly extract patches from the images. For the nematode dataset, 3000 positive and 3000 negative patches of size 1504 x 2056 are extracted from 3300 images. The rest 667 images are left as test. Similarly, we crop 1172 positive and 4854 negative examples of size 384x512 from 387 cyst images, leaving the rest 100 images as test.

All images were resized to 384x512, and normalized to 0 mean and 1 standard deviation for input. An U-Net [16] was used as the segmentation backbone. The RMSprop optimizer was used for training with initial learning rate 1e-4. All experiment runs followed the same training routine: (1) train for 100 epoches for the classification training and 20 epoches for each refinement training iteration, (2) set aside 10% training images for validation purpose and (3) save the best model with respect to the validation loss.

C. Experiments and results

For quantitative evaluation, we report the classification accuracy and segmentation Dice coefficient. To make the score more insightful, we only consider images containing objects when computing the Dice coefficient. Since the classification accuracy is high, if we take correctly classified negative image as Dice 1, the overall Dice coefficient will distortedly high. The results are summarized in Table I.

In terms of classification accuracy, there is no statistical difference observed for different pooling methods, except the case where GAP is used. We argue that performance reduction roots in the fact that a sufficient number of pixel must be activated to make the classification score high enough. And this is not naturally easy for images with sparsely distributed small object. The results that performance loss is greater on the nematode dataset proves this analysis from another aspect.

Even the classification task is well trained, not all activation maps can be used as meaningful segmentation or further refined. GMP activates only a few pixels in the object boundary area (Figure 2), which is a very limited indication of the position and extent of objects. In the same situation but for different reasons, GAP aggressively highlights a much larger region containing objects of interest, sometimes even close to covering the entire image.

TABLE I

QUANTITATIVE EVALUATION. THE DICE COEFFICIENT IS REPORTED AS THE MAIN EVALUATION SCORE FOR DIFFERENT POOLING METHODS AND DIFFERENT REFINEMENT STEP.

	Dice (%)	GAP	GWP1	GWP3	GWP5	GWP7	GWP9	GWP11	GWP13	GMP
Nematode	Raw Seg.	2.29	9.31	14.81	24.10	33.64	**46.56**	35.02	20.27	5.83
	Superpixel Voting	-	51.14	56.04	61.22	59.55	**62.94**	28.90	30.69	4.23
	Refine Train Ite. 1	-	66.06	72.71	74.93	74.86	**74.95**	34.89	36.51	-
	Refine Train Ite. 2	-	70.51	75.60	78.72	**79.72**	77.90	35.07	41.12	-
	Cls. Acc. (%)	76.66	96.12	96.20	95.96	95.88	96.52	95.72	96.77	95.32
Cyst	Raw Seg.	4.82	21.70	33.51	44.01	**47.48**	28.46	14.67	8.64	2.87
	Superpixel Voting	-	27.84	36.34	**52.67**	51.79	44.65	19.89	3.65	0.66
	Refine Train Ite. 1	-	32.40	39.60	54.67	**55.87**	47.74	27.79	0.85	-
	Refine Train Ite. 2	-	31.71	38.95	**58.51**	56.10	51.79	30.43	-	-
	Cls. Acc. (%)	86.23	90.88	91.50	88.84	89.95	89.24	90.91	90.52	89.33

By contrast, our proposed GWP can adapt well to the object extent in a large value range of r. Although the Dice coefficient of raw activation maps under different r values varies in a relatively large range, all provide a valid starting point for further refinement. Both superpixel voting and refinement training consistently improve the segmentation quality. Averaged on all GWP experiments, superpixel voting improves the Dice coefficient from 26.24% to 50.07% on the nematode dataset, and from 28.35% to 33.83% on the cyst dataset. Training supervised by the refined segmentation further promotes the performance by 22.54% and 29.24% after the first and second iteration on the nematode dataset, while the improvement is 9.34% and %12.96 on the cyst dataset. The improvement saturates after 2-3 refinement iterations.

On the nematode dataset, GWP with $r = 7$ achieves the best Dice coefficient 79.72%. The best cyst segmentation is 58.51% Dice coefficient at $r = 5$. As a baseline, we also have trained the segmentation model fully supervised with the pixel-level ground truth, which generates 89.76% and 72.22% Dice coefficient on the nematode and cyst dataset, respectively. Although there is still a performance gap with pixel-wise supervision, our method gives quite promising results with minimal annotation effort, which is a valid segmentation for most analysis, such as quantification and phenotyping.

IV. CONCLUSION AND OUTLOOK

We have proposed an approach for training segmentation models with image-level supervision, explicitly considering the segmentation of fine structures. Pooling features using GWP and supervised by the classification loss, the model can coarsely locate target objects. Through superpixel voting and refinement training, the segmentation is significantly refined and achieves at best 79.72% and 58.51% Dice coefficient on the nematode and cyst dataset, respectively.

Although negative examples, as employed in our training scheme, are often easy to collect, public segmentation datasets tend to contain only images with objects. Here, we have performed evaluation on nematode datasets in a agricultural application scenario. Further experiments should be conducted on other suitable datasets. In addition, extension to multiple classes is straightforward in principle, but the actual performance needs to be studied experimentally.

REFERENCES

[1] H. Qu, P. Wu, Q. Huang, J. Yi, G. M. Riedlinger, S. De, D. N. Metaxas, "Weakly Supervised Deep Nuclei Segmentation using Points Annotation in Histopathology Images," in MICCAI, Shenzhen, China, 2019, pp. 390-400.

[2] Z. Ji, Y. Shen, C. Ma, M. Gao, "Scribble-Based Hierarchical Weakly Supervised Learning for Brain Tumor Segmentation," in MICCAI, Shenzhen, China, 2019, pp. 175-183.

[3] L. Yang, Y. Zhang, Z. Zhao, H. Zheng, P. Liang, M. T. C. Ying, A. Ahuja, D. Chen, "BoxNet: Deep Learning Based Biomedical Image Segmentation Using Boxes Only Annotation," ArXiv, 2018.

[4] Q. Li, A. Arnab, P. H. S. Torr, "Weakly- and Semi-Supervised Panoptic Segmentation," in ECCV, Munich, Germany, 2018, pp. 102-118.

[5] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, T. S. Huang, "Revisiting Dilated Convolution: A Simple Approach for Weakly- and Semi-Supervised Semantic Segmentation," in CVPR, Salt Lake City, USA, 2018, pp. 7268-7277.

[6] A. Kolesnikov, C. H. Lampert, "Seed, Expand and Constrain: Three Principles for Weakly-Supervised Image Segmentation," in ECCV, Amsterdam, Netherlands, 2016, pp. 695-711.

[7] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," in CVPR, Las Vegas, USA, 2016, pp. 2921-2929.

[8] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in ICCV, Venice, Italy, 2017, pp. 618-626.

[9] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen and S. Sclaroff, "Top-Down Neural Attention by Excitation Backprop," International Journal of Computer Vision, vol 126, no. 10, pp. 1084-1102, Oct 2018.

[10] P. Krähenbühl, V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in NIPS, Granada, Spain, pp. 109-118.

[11] T. H. Askary, "Limitations, research needs and future prospects in the biological control of phytonematodes", in Biocontrol Agents of Phytonematodes, T. H. Askary, P. R. P. Martinelli, Wallingford, UK: CABI, 2015, pp. 446–454.

[12] P.F. Felzenszwalb, D. P. Huttenlocher, "Efficient graph-based image segmentation," International Journal of Computer Vision, vol. 59, no. 2, pp. 167-181, Sep 2004.

[13] M. M. Rahaman, D. Chen, Z. Gillani, C. Klukas, M. Chen, "Advanced phenotyping and phenotype data analysis for the study of plant growth and development," Frontiers in Plant Science, vol. 6, Aug 2015.

[14] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in DLMIA, Québec City, Canada, pp. 240–248.

[15] E. N. Mortensen, W. A. Barrett, "Intelligent Scissors for Image Composition," in SIGGRAPH, New York, USA, pp. 191–198.

[16] O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in MICCAI, Munich, Germany, pp. 234-241.