SCRIBBLE-SUPERVISED CELL SEGMENTATION USING MULTISCALE CONTRASTIVE REGULARIZATION

Hyun-Jic Oh [1]   *Kanggeun Lee* [2]   *Won-Ki Jeong* [1,*]

[1] Department of Computer Science and Engineering, Korea University, Seoul, Korea
[2] School of Electrical and Computer Engineering, UNIST, Ulsan, Korea

ABSTRACT

Current state-of-the-art supervised deep learning-based segmentation approaches have demonstrated superior performance in medical image segmentation tasks. However, such supervised approaches require fully annotated pixel-level ground-truth labels, which are labor-intensive and time-consuming to acquire. Recently, *Scribble2Label* (S2L) demonstrated that using only a handful of scribbles with self-supervised learning can generate accurate segmentation results without full annotation. However, owing to the relatively small size of scribbles, the model is prone to overfit and the results may be biased to the selection of scribbles. In this work, we address this issue by employing a novel multiscale contrastive regularization term for S2L. The main idea is to extract features from intermediate layers of the neural network for contrastive loss so that structures at various scales can be effectively separated. To verify the efficacy of our method, we conducted ablation studies on well-known datasets, such as Data Science Bowl 2018 and MoNuSeg. The results show that the proposed multiscale contrastive loss is effective in improving the performance of S2L, which is comparable to that of the supervised learning segmentation method.

Index Terms— Contrastive Learning, Weakly Supervised Learning, Cell Segmentation

1. INTRODUCTION

Cell segmentation [1] is a fundamental and important image processing method that is necessary in many biomedical image analysis tasks, such as cell counting and cell morphology analysis. Supervised learning with convolutional neural networks (CNNs) has shown superior performance in cell segmentation, but it requires a sufficient number of manually annotated pixel-level labels for robust results, which demands a labor-intensive and time-consuming label generation process. In addition, even the labels generated by experts are not perfect and may contain errors, especially in complex or subtle boundaries, which impairs the performance of supervised learning-based approaches.
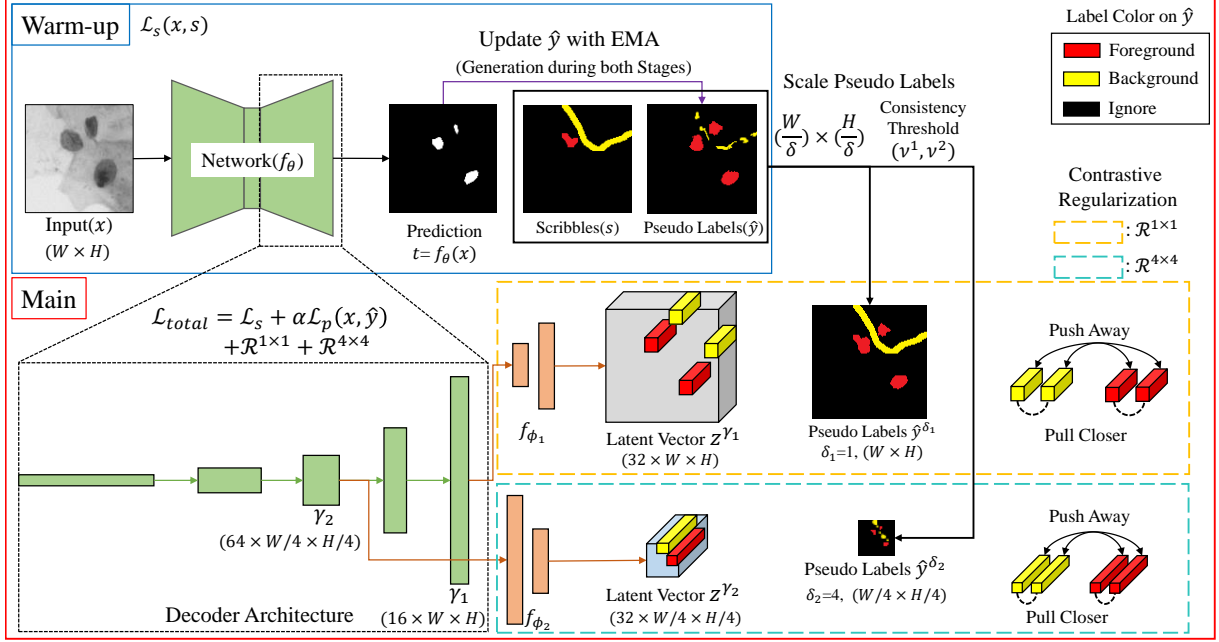
To alleviate the need for manual annotations, various weakly supervised segmentation algorithms that do not use full ground-truth labels have been introduced. Bearman *et al.* [2] proposed a weakly supervised segmentation scheme that learns from point annotations and generic objectness prior [3], providing the guidelines for separating foreground and background. More recently, learning from point annotations [4, 5, 6] in the medical image domain showed the potential of weakly supervised learning, which closes the performance gap to fully supervised learning. Another approach is to use sparse scribbles rather than per-object point annotation. Lin *et al.* [7] used a super-pixel-based pseudo-label generated by scribble annotations. Scribble2Label (S2L) [8] proposed scribble-based supervised learning with pseudo-labeling on an unlabeled area. Even though S2L achieved robust results, it still has a limitation of overfitting scribble annotations, which may impair the segmentation performance.

The main motivation of this work is to address the limitation of S2L by introducing a new regularization energy, which stems from the observation of recent advances in contrastive learning. Recently, self-supervised contrastive learning (SimCLR) [9] has shown notable performance as a significant variant of self-supervised learning. The main intuition behind SimCLR is that transformation of similar images should have similar representations (i.e., latent features of a neural network), while different images should have different representations in the feature space. Khosla *et al.* [10] demonstrated the possibility of the scalability of self-supervised contrastive learning to supervised contrastive learning in an image classification task. Supervised contrastive learning aims to push the anchor far away from samples in different classes, and conversely, to pull the anchor close to samples in the same class. However, existing contrastive learning works mostly focus on image classification. Contrastive learning in image segmentation is in its early stage, and only a few works have been attempted on image segmentation [11, 12].

In this study, we propose multiscale contrastive regularization based on pseudo supervision produced by scribble-supervised learning to improve the performance of cell segmentation. The main contributions of our work are summarized as follows:

1. We introduce a novel adaptation of supervised contrastive

---

*Corresponding author: wkjeong@korea.ac.kr

Fig. 1. An overview of the proposed method. Because our method is based on S2L, we briefly described it [8] with the proposed method. First, the warm-up stage proceeds scribble-supervised learning and pseudo label generation. Next, while continuing the same process on main stage, we additionally conducts pseudo-label based learning and multiscale contrastive regularization. Putting each output from decoder layer γ into projection head f_ϕ, we obtain latent vector z^γ. Here, we use scribbles and the generated pseudo labels to annotate the label of the latent vector. According to the labeling standard ν and the resolution factor δ, $\hat{y}^\delta$ is generated. With the assigned labels and latent vectors, we compute contrastive loss.

learning from image classification into a weakly supervised pixel-level image segmentation.

2. We introduce a generalized multiscale contrastive regularization applicable to weakly supervised learning by employing supervised contrastive regularization on various image resolutions.

3. We demonstrate the efficacy of the proposed method by extending S2L and performing quantitative and qualitative analyses on well-known datasets (e.g., Data Science Bowl 2018 and MoNuSeg).

2. METHOD

2.1. S2L Overview

The original S2L [8] consists of two stages: warm-up and main stages. As described in Fig. 1, in the warm-up stage, the CNN only learns from a set of scribbled pixels Ω, and periodically generates pseudo labels through a filtering method based on the consistency threshold and exponential moving average (EMA) method. We rewrote the scribble-based supervised binary cross-entropy loss $\mathcal{L}_s$ [8] as follow:

$$\mathcal{L}_s(x, s) = \frac{1}{|\Omega|} \sum_{i \in \Omega} \text{BCE}(s_i, t_i) \quad (1)$$

where x is an input image, $t = f_\theta(x)$ is a vectorized prediction generated by an arbitrary CNN f_θ, and s_i is a vectorized scribble label indexed by $i \in \Omega$. Second, in the main stage, they [8] retrain the function f_θ with pseudo-labels [8] through the following main loss:
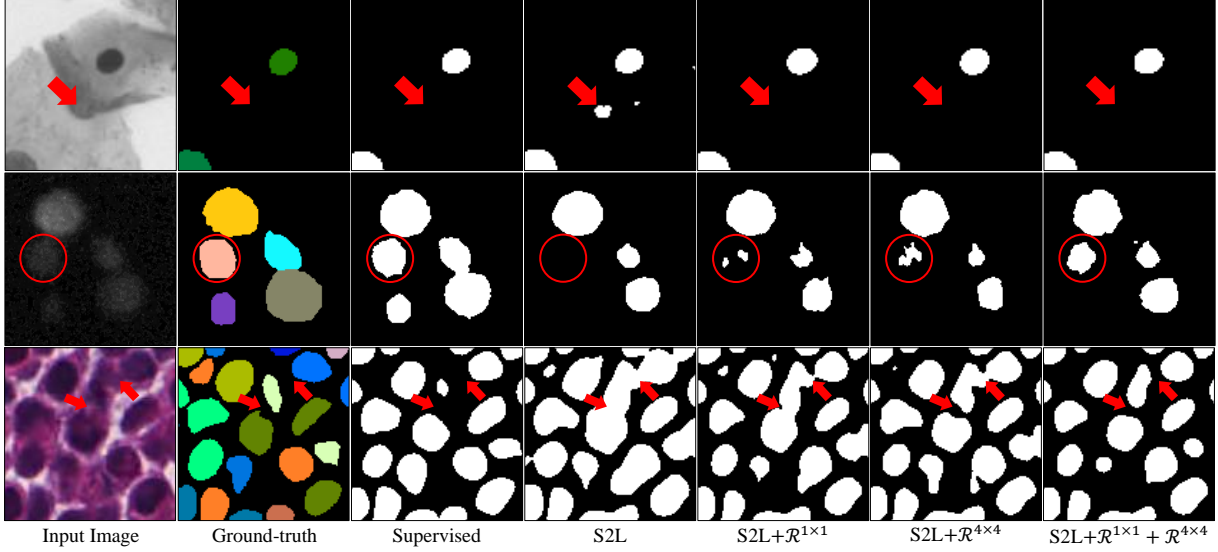
$$\mathcal{L}_{\text{S2L}}(x, s, \hat{y}) = \mathcal{L}_s(x, s) + \alpha \mathcal{L}_p(x, \hat{y}) \quad (2)$$

$$\mathcal{L}_p(x, \hat{y}) = \frac{1}{|\Omega_p|} \sum_{i \in \Omega_p} \text{BCE}(\hat{y}_i, t_i) \quad (3)$$

where $\hat{y}$ is a vectorized pseudo-label and Ω_p is a set of pseudo-labels. We introduce the modified main stage in the next step as depicted in Fig. 1.

2.2. Multiscale Contrastive Regularization

This section introduces multiscale contrastive regularization. We reformulate the main loss in Eq. 2 to employ multiscale contrastive regularization that restricts the convergence space of the arbitrary CNN f. Let $f_\theta(x)$ be composed of a sequence of transformations $f^{n-1} \circ ... \circ f^1 \circ f^0(x)$. We select a latent feature vector $v^\gamma = f^\gamma \circ ... \circ f^0(x)$ such that $0 < \gamma \leq n$, and pass through a projection head similar to [9] as $f_\phi(v^\gamma) = z^\gamma$. The projection head $f_\phi : \mathbb{R}^{C \times (HW/\delta^2)} \to \mathbb{R}^{C' \times (HW/\delta^2)}$ consists of combinations of convolutional layers, batch normalization layers, and activation functions. Now, we can define

Fig. 2. Qualitative comparison between ground-truth, supervised, and our method (with various loss combinations). Top row: DSB-BF, middle row: DSB-Fluo [13], and bottom row: MoNuSeg [14] datasets. For the input image of DSB-Fluo data, we increased the image contrast for better visualization. Each ground truth image includes per-instance labels. Some notable regions are marked with arrows and circles.

pseudo-supervised multi-scale contrastive regularization with respect to γ and δ as:

$$\mathcal{L}_c(\hat{y}, v; f_\phi, \gamma, \delta, \tau) =$$
$$\frac{\delta^4}{W^2 H^2} \sum_i \sum_j \text{BCE}(\hat{y}_i^\delta \wedge \hat{y}_j^\delta, \sigma(\frac{z_i^\gamma}{||z_i^\gamma||_2} \cdot \frac{z_j^\gamma}{||z_j^\gamma||_2} / \tau)) \quad (4)$$

where σ is a sigmoid function, $\cdot$ denotes the dot product, and $\wedge$ symbol denotes element-wise "or" logical operator. And,

$$\hat{y}_i^\delta = \begin{cases} 0, & \text{if } \frac{1}{\delta^2} ||\hat{y}_{i:i+\delta^2}||_1 \leq \nu_1 \text{ or } i \in \Omega^- \\ 1, & \text{if } \frac{1}{\delta^2} ||\hat{y}_{i:i+\delta^2}||_1 \geq \nu_2 \text{ or } i \in \Omega^+ \quad (5) \\ \text{Ignore}, & \text{otherwise} \end{cases}$$

In other words, $\hat{y}^\delta$ is a down-scaled vector of a pseudo-label generated by the factor $\delta \in \{2^i | i \in \mathbb{N}_0\}$ with consistency threshold values ν_1 and ν_2 for contrastive regularization at various resolutions. $i \in \Omega^-$ and $i \in \Omega^+$ denote the background and foreground scribble labels, respectively. In supervised contrastive learning, the softmax function can control a single positive sample with an anchor; however, it is an inefficient regularization in the segmentation task because of the large number of positive pixels. Unlike the supervised contrastive regularization study [10], we employ a logistic function to scale the element-wise probability of pseudo-labels instead of a softmax function. The logistic function makes it feasible to handle multiple positive pixels for each sample x simultaneously. Now, we can utilize supervised contrastive regularization in the latent space of various resolution features

(i.e., it is extracted features on intermediate layers of encoder-decoder-like shaped model). Consequently, we can define the main loss function as a combination of the main loss of S2L with multiscale contrastive regularization as follows:

$$\mathcal{L}_{total}(x, s, \hat{y}, v; f_\phi, \gamma, \delta, \tau) = \mathcal{L}_{\text{S2L}}(x, s, \hat{y})$$
$$+ \lambda_1 \mathcal{L}_c(\hat{y}, v; f_{\phi_1}, \gamma_1, \delta_1, \tau_1)$$
$$+ \lambda_2 \mathcal{L}_c(\hat{y}, v; f_{\phi_2}, \gamma_2, \delta_2, \tau_2) \quad (6)$$

where α, λ_1, and λ_2 are weight parameters. Although Eq. 4 can be extended to multiple regularization terms for all possible δ, we empirically chose two regularization terms with different scale factors $\delta_1 = 1$ and $\delta_2 = 4$ in our experiment to balance between the training efficiency and the performance.

3. EXPERIMENT

3.1. Data

We used three different cell image datasets. The first two sets of nuclei 2D images were obtained from BBBC038v1 [13] known as Data Science Bowl (DSB) 2018 data. In the DSB dataset, we used the Stage 1 training dataset with fully annotated masks, which were 16 bright-field (DSB-BF) 1000×1000 images and 542 fluorescence (DSB-Fluo) images of various sizes. We conducted 4-fold cross validation on these two datasets, so we splits each dataset into 60% for training set, 20% for the validation set, and 20% for the test set. The third dataset was MoNuSeg [14], which consisted of 30 training set and 14 test set 1000×1000 histopathology

Table 1. Quantitative experiment results representing accuracy in the format of mDice / IoU. The first- and second-best results are highlighted in red and blue, respectively.

Label	Method	DSB-BF	DSB-Fluo	MoNuSeg
Scribble	S2L[8]	0.6980 / 0.8108	0.8235 / 0.8419	0.7098 / 0.6586
	S2L+$\mathcal{R}^{1\times1}$	0.7068 / 0.8153	0.8267 / 0.8565	0.7165 / 0.6715
	S2L+$\mathcal{R}^{4\times4}$	0.7078 / 0.8292	0.8234 / 0.8505	0.7205 / 0.6603
	S2L+$\mathcal{R}^{1\times1}$+$\mathcal{R}^{4\times4}$	0.7199 / 0.8368	0.8268 / 0.8570	0.7128 / 0.6639
Full	Supervised	0.7305 / 0.8665	0.8204 / 0.8790	0.7646 / 0.6867

images. We conducted 5-fold cross validation and made predictions (inference) on the test set images. To manually draw scribbles, we referenced fully annotated masks.

3.2. Implementation Details

Our baseline network is U-Net [15] with five layers of decoder blocks, using the ResNet-34 [16] backbone for the experiment. We used constant values of $\alpha = 0.5$, $\lambda_1 = 0.5$, and $\lambda_2 = 10$. For multiscale contrastive regularization, we empirically chose the third and last layers of the decoder (i.e., γ_1 and γ_2). Let us denote $\lambda_1\mathcal{L}_c$ as $\mathcal{R}^{1\times1}$ and $\lambda_2\mathcal{L}_c$ as $\mathcal{R}^{4\times4}$. Because the loss operates more tolerant to similar latent vectors for the larger temperature parameter τ on contrastive loss [17], we set the temperature $\tau_1 = 0.3$ and $\tau_2 = 0.1$. For the latent vector label allocation, we set $\nu_1 = 0$ and $\nu_2 = 1$ to use more reliable pseudo labels. To reduce the cost of contrastive loss computation, we randomly sampled from each scale per class, setting 6000 as the maximum pixel number per class in pixel sampling. We used the same hyper-parameters used in the original S2L for Eq. 2.

3.3. Results

For performance assessment, we compared our results with those of S2L and a supervised learning method using full annotation. Additionally, we conducted an ablation study to describe the segmentation performance of each regularization for a single resolution.

For quantitative comparisons, we used the intersection of union (IoU) to evaluate the overall (semantic) segmentation accuracy and mean of the Dice-coefficient (mDice) [18] to measure the per-instance segmentation accuracy (see Table 1). In DSB-BF and DSB-Fluo cases, multiscale contrastive regularization outperformed S2L and single-scale contrastive regularization. In the case of DSB-BF, single-scale contrastive regularization with $\mathcal{R}^{4\times4}$ showed the second highest performance. Because $\mathcal{R}^{4\times4}$ has a larger field-of-view, it has the advantage of separating regions that have similar intensity but structurally different. In DSB-Fluo, $\mathcal{R}^{1\times1}$ outperformed $\mathcal{R}^{4\times4}$ and showed a similar performance as

the combination of $\mathcal{R}^{1\times1}$ and $\mathcal{R}^{4\times4}$. $\mathcal{R}^{1\times1}$ is more effective in distinguishing subtle intensity differences by comparing pixels in the highest resolution. For the MoNuSeg data, $\mathcal{R}^{4\times4}$ performed best.

Fig. 2 shows the qualitative results with input, ground-truth, and the proposed method using various experimental settings. In the case of DSB-BF, S2L falsely predicted the nonnuclear region as the nucleus (see the arrows in the first row of Fig. 2) that has a similar intensity as the nucleus. We believe this is owing to overfitting because scribbles are drawn inside cells, so the intensity is a dominant factor for the decision rather than the shape. However, contrastive regularization distinguished the cell region correctly because different shapes could be separated by feature similarity across different scales. In the case of DSB-Fluo, our proposed method detected the cell regions where S2L failed to detect. Note that our two scale-based regularization method (i.e., combining $\mathcal{R}^{1\times1}$ and $\mathcal{R}^{4\times4}$) detected the cell shape and boundary area that the other approaches failed to predict. In the case of MoNuSeg, S2L failed to separate cells nearby, which was improved by employing contrastive regularization.

4. CONCLUSION

In this study, we introduced supervised contrastive learning with multiscale features in scribble-based weakly supervised learning for cell segmentation. We demonstrated that two contrastive regularizers on different scales can achieve superior performance compared to S2L in DSB-BF and DSB-Fluo datasets, and a single-scale contrastive regularizer can also boost the results on the MoNuSeg dataset. We conclude that the proposed contrastive regularization is effective in avoiding the overfitting issue of scribble-based weakly supervised learning.

In future works, we plan to extend the current framework to adaptive contrastive regularization, which can leverage all the latent features from every resolution. Applying the proposed method to different image modalities, such as natural images, would be another interesting future research direction.

5. COMPLIANCE WITH ETHICAL STANDARDS

6. ACKNOWLEDGMENTS

7. REFERENCES

[1] E. Meijering, "Cell segmentation: 50 years down the road [life sciences]," *IEEE Signal Processing Magazine*, vol. 29, pp. 140–145, August 2012.

[2] A. Bearman, O. Russakovsky, V. Ferrari, and Li Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *Eur. Conf. Comput. Vis.* Springer, 2016, vol. 9911, pp. 549–565.

[3] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 34, no. 11, pp. 2189–2202, 2012.

[4] I. Yoo, D. Yoo, and K. Paeng, "Pseudoedgenet: Nuclei segmentation only with point annotations," in *Proc. Int. Conf. Med. Image Comput. and Comp.-Assist. Intervent.* Springer, 2019, vol. 11764, pp. 731–739.

[5] Z. Chen, Z. Chen, J. Liu, Q. Zheng, Y. Zhu, Y. Zuo, Z. Wang, X. Guan, Y. Wang, and Y. Li, "Weakly supervised histopathology image segmentation with sparse point annotations," *IEEE Journal of Biomed. and Health Inform.*, vol. 25, pp. 1673–1685, May 2021.

[6] T. Zhao and Z. Yin, "Weakly supervised cell segmentation by point annotation," *IEEE Trans. Med. Imag.*, vol. 40, pp. 2736–2747, December 2021.

[7] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recoginit.*, 2016, pp. 3159–3167.

[8] H. Lee and W. K. Jeong, "Scribble2label: Scribble-supervised cell segmentation via self-generating pseudo-labels with consistency," in *Proc. Int. Conf. Med. Image Comput. and Computer-Assisted Intervention*. Springer, 2020, vol. 12261, pp. 14–23.

[9] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, vol. 119, pp. 1597–1607.

[10] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," 2020, vol. 33, pp. 1–23.

[11] Y. Zhong, B. Yuan, H. Wu, Z. Yuan, J. Peng, and Y. Wang, "Pixel contrastive-consistent semi-supervised semantic segmentation," 2021, pp. 7273–7282.

[12] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," in *Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 12546–12558.

[13] J. Caicedo, A. Goodman, K. Karhohs, B. Cimini, J. Ackerman, M. Haghighi, C. Heng, T. Becker, M. Doan, C. McQuin, M. Rohban, S. Singh, and A. Carpenter, "Nucleus segmentation across imaging experiments: the 2018 data science bowl," *Nature Methods*, vol. 16, pp. 1247–1253, Dec 2019.

[14] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi, "A dataset and a technique for generalized nuclear segmentation for computational pathology," *IEEE Trans. Med. Imag.*, vol. 36, pp. 1550–1560, March 2017.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015, vol. 9351, pp. 234–241.

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

[17] F. Wang and H. Liu, "Understanding the behaviour of contrastive loss," 2021, pp. 2495–2504.

[18] K. Nishimura, D. Ker, and R. Bise, "Weakly supervised cell instance segmentation by propagating from detection response," 2019, vol. 11764, pp. 649–657.