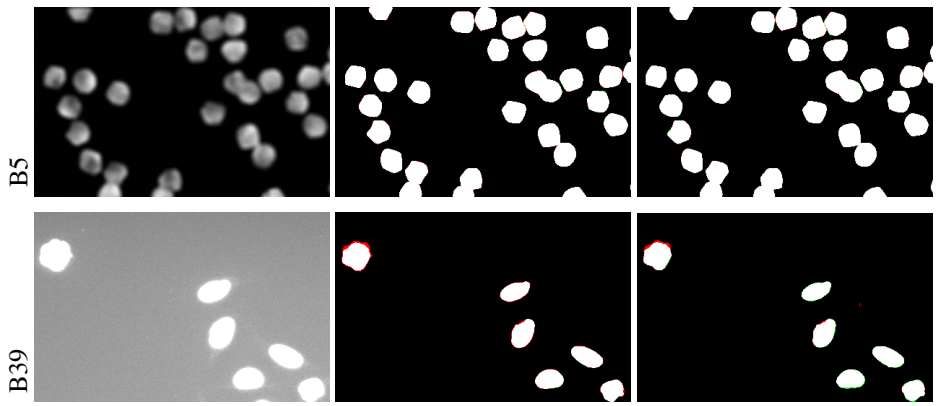
# Supplementary Material - Knowing What to Label for Few Shot Microscopy Image Cell Segmentation

Youssef Dawoud [1*] Arij Bouazizi [1,2] Katharina Ernst [3] Gustavo Carneiro [4]
Vasileios Belagiannis [5]

[1] Universität Ulm, Ulm, Germany
[2] Mercedes-Benz AG, Stuttgart, Germany
3 Ulm University Medical Center, Ulm, Germany
4 Centre for Vision, Speech and Signal Processing, University of Surrey, United Kingdom
5 Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
youssef.dawoud@uni-ulm.de

In the supplementary material we present the Wilcoxon significance test of our approach against random, MC-dropout, and entropy baselines in Tab. 1. Our null hypothesis ($H_0$) assumes that the mIoU of the baseline is greater than our approach i.e. $\text{mIoU}_{baseline} > \text{mIoU}_{Ours}$. On the other hand, the alternate ($H_a$) claims the opposite i.e. $\text{mIoU}_{baseline} < \text{mIoU}_{Ours}$. We use significance level of 0.05. The results clearly demonstrate the significance of our approach in many cases (36 out of 75) relative to the baselines which supports the alternate hypothesis. We show additional visual results for B5 and B39 datasets and their corresponding numerical results in Tab. 7. Furthermore, we report the numerical results of our ablation studies in Tab. 4 to 3. First, in Tab. 4 and 5 we show the impact of limiting the image patches extracted per target image for ssTEM and EM target datasets respectively. Second, we present the impact of increasing the number of patches extracted per 1-shot for ssTEM target dataset in Tab. 4. Afterwards, in Tab. 2 we report the performance on target datasets TNBC, EM, and ssTEM between using the pretrained model and the model trained using pseudo-label segmentation learning for support set selection and fine-tuning. In Tab. 3 we report the results of using pixel-level augmentations compared to affine augmentations. Finally, we report the fine-tuning performance of human-expert selection compared to our selection approach in Tab. 6.




(a) Input (b) Random (c) Ours

Figure 1. Visual Result. We visually compare our scoring function (Ours) to random selection using the FCRN architecture at $|\mathcal{B}| = 3$-shots. The red colour corresponds to false positive, the green colour to false negative, the black colour to true negative, and the white colour to true positive. Best viewed in colour.

| Target: TNBC | | | | | |
|---|---|---|---|---|---|
| Method | 1-shot ($\|\mathcal{B}\| = 100$) | 3-shot ($\|\mathcal{B}\| = 300$) | 5-shot ($\|\mathcal{B}\| = 500$) | 7-shot ($\|\mathcal{B}\| = 700$) | 10-shot ($\|\mathcal{B}\| = 1000$) |
| Entropy | 0.003 | 0.023 | 0.062 | 0.037 | 0.023 |
| MC-dropout | 0.003 | 0.023 | 0.063 | 0.018 | 0.023 |
| Random | 0.008 | 0.011 | 0.011 | 0.037 | 0.003 |
| Target: EM | | | | | |
| Method | 1-shot ($\|\mathcal{B}\| = 400$) | 3-shot ($\|\mathcal{B}\| = 1200$) | 5-shot ($\|\mathcal{B}\| = 2000$) | 7-shot ($\|\mathcal{B}\| = 2800$) | 10-shot ($\|\mathcal{B}\| = 4000$) |
| Entropy | 0.25 | 0.003 | 0.046 | 0.006 | 0.14 |
| MC-dropout | 0.29 | 0.003 | 0.008 | 0.003 | 0.003 |
| Random | 0.08 | 0.005 | 0.057 | 0.057 | 0.029 |
| Target: ssTEM | | | | | |
| Method | 1-shot ($\|\mathcal{B}\| = 500$) | 3-shot ($\|\mathcal{B}\| = 1500$) | 5-shot ($\|\mathcal{B}\| = 2500$) | 7-shot ($\|\mathcal{B}\| = 3200$) | 10-shot ($\|\mathcal{B}\| = 5000$) |
| Entropy | 0.037 | 0.08 | 0.96 | 0.89 | 0.36 |
| MC-dropout | 0.03 | 0.12 | 0.81 | 0.94 | 0.71 |
| Random | 0.011 | 0.006 | 0.14 | 0.52 | 0.084 |
| Target: B39 | | | | | |
| Method | 1-shot ($\|\mathcal{B}\| = 100$) | 3-shot ($\|\mathcal{B}\| = 300$) | 5-shot ($\|\mathcal{B}\| = 500$) | 7-shot ($\|\mathcal{B}\| = 700$) | 10-shot ($\|\mathcal{B}\| = 1000$) |
| Entropy | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 |
| MC-dropout | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 |
| Random | 0.99 | 0.98 | 0.99 | 0.99 | 0.99 |
| Target: B5 | | | | | |
| Method | 1-shot ($\|\mathcal{B}\| = 100$) | 3-shot ($\|\mathcal{B}\| = 300$) | 5-shot ($\|\mathcal{B}\| = 500$) | 7-shot ($\|\mathcal{B}\| = 700$) | 10-shot ($\|\mathcal{B}\| = 1000$) |
| Entropy | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| MC-dropout | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| Random | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |

Table 1. Wilcoxon signed-rank significance test results of Ours vs $\{$ Entropy, MC-dropout, Random $\}$. Our null hypothesis $H_0$ assumes that the mIoU of the baseline is greater than our approach. While the alternate hypothesis $H_a$ claims the opposite. We set the significance level at 0.05 i.e. we reject $H_0$ if $p < 0.05$ . We report the $p$-values for targets TNBC, EM, ssTEM, B39, and B5.

| Target: TNBC | | | | | |
|---|---|---|---|---|---|
| Model | 1-shot | 3-shot | 5-shot | 7-shot | 10-shot |
| $f_\theta(\mathbf{x})$ | 44.5% ±5.0 | 47.8% ±6.2 | 47.7% ±7.5 | 48.3% ±7.0 | 48.7% ±6.2 |
| $f_{\theta'}(\mathbf{x})$ | 47.1% ±3.5 | 47.8% ±5.8 | 47.7% ±5.0 | 48.0% ±5.2 | 49.2% ±4.9 |
| Target: EM | | | | | |
| Model | 1-shot | 3-shot | 5-shot | 7-shot | 10-shot |
| $f_\theta(\mathbf{x})$ | 52.7% ±1.9 | 64.8% ±1.7 | 66.4% ±2.0 | 68.0% ±2.0 | 70.4% ±1.0 |
| $f_{\theta'}(\mathbf{x})$ | 62.0% ±3.2 | 69.8% ±2.7 | 70.6% ±3.3 | 73.1% ±2.9 | 73.7% ±3.2 |
| Target: ssTEM | | | | | |
| Model | 1-shot | 3-shot | 5-shot | 7-shot | 10-shot |
| $f_\theta(\mathbf{x})$ | 50.6% ±2.8 | 62.8% ±3.0 | 64.8% ±1.7 | 66.7% ±3.4 | 67.2% ±3.0 |
| $f_{\theta'}(\mathbf{x})$ | 51.3% ±3.2 | 63.3% ±3.2 | 64.2% ±3.2 | 67.3% ±2.9 | 68.7% ±2.6 |

Table 2. Effect on mIoU results of the pseudo-label segmentation learning and selection on the target data sets TNBC, EM, and ssTEM.

| Target: TNBC | | | | | |
|---|---|---|---|---|---|
| Augmentation | 1-shot | 3-shot | 5-shot | 7-shot | 10-shot |
| $A = \{R, T_x, T_y\}$ | 40.6% ±4.2 | 41.4% ±4.6 | 44.7% ±5.0 | 46.9% ±4.2 | 48.7% ±3.5 |
| $A = \{C, B, S\}$ | 47.1% ±3.5 | 47.8% ±5.8 | 47.7% ±5.0 | 48.0% ±5.2 | 49.2% ±4.9 |
| Target: EM | | | | | |
| Augmentation | 1-shot | 3-shot | 5-shot | 7-shot | 10-shot |
| $A = \{R, T_x, T_y\}$ | 47.8% ±2.6 | 55.5% ±1.7 | 60.4% ±3.6 | 64.0% ±3.2 | 67.3% ±2.7 |
| $A = \{C, B, S\}$ | 62.0% ±3.2 | 69.8% ±2.7 | 70.6% ±3.3 | 73.1% ±2.9 | 73.7% ±3.2 |
| Target: ssTEM | | | | | |
| Augmentation | 1-shot | 3-shot | 5-shot | 7-shot | 10-shot |
| $A = \{R, T_x, T_y\}$ | 49.4% ±4.4 | 61.9% ±3.4 | 64.2% ±4.0 | 67.6% ±2.8 | 69.7% ±3.0 |
| $A = \{C, B, S\}$ | 51.3% ±3.2 | 63.3% ±3.2 | 64.2% ±3.2 | 67.3% ±2.9 | 68.7% ±2.6 |

Table 3. Effect on mIoU results of the data augmentation and selection on the target data set TNBC, EM, and ssTEM.

| Method | 1-shot ($\lvert\mathcal{B}\rvert = 100$) | 3-shot ($\lvert\mathcal{B}\rvert = 300$) | 5-shot ($\lvert\mathcal{B}\rvert = 500$) | 7-shot ($\lvert\mathcal{B}\rvert = 700$) | 10-shot ($\lvert\mathcal{B}\rvert = 1000$) |
|---|---|---|---|---|---|
| Entropy | 17.3%±1.3 | 38.5% ±3.0 | 49.5% ±4.0 | 54.2% ±3.7 | 58.7% ±3.3 |
| MC-dropout | 17.2%±1.3 | 38.8% ±2.7 | 49.4% ±3.5 | 54.1% ±4.0 | 58.8% ±3.1 |
| Random | 20.6% ±2.0 | 39.1% ±3.7 | 47.2% ±3.7 | 52.1% ±4.5 | 56.1% ±4.5 |
| ***Ours*** | **25.2% ±2.0** | **41.8% ±2.9** | **52.3% ±2.4** | **55.4% ±3.0** | **60.4% ±2.2** |
| 1-shot | | | | | |
| | $\lvert\mathcal{B}\rvert = 100$ | $\lvert\mathcal{B}\rvert = 200$ | $\lvert\mathcal{B}\rvert = 300$ | $\lvert\mathcal{B}\rvert = 400$ | $\lvert\mathcal{B}\rvert = 500$ |
| Entropy | 17.3% ±1.3 | 29.9% ±2.4 | 38.4% ±3.0 | 45.9% ±3.4 | 48.2% ±3.2 |
| MC-dropout | 17.2% ±2.0 | 29.8% ±2.4 | 38.8% ±2.7 | 45.5% ±3.2 | 49.3% ±3.4 |
| Random | 20.6% ±2.0 | 31.1% ±3.1 | 38.5% ±3.6 | 45.1% ±4.5 | 47.0% ±4.8 |
| ***Ours*** | **25.2% ±2.0** | **35.1% ±2.6** | **41.8% ±3.4** | **48.8% ±2.7** | **51.3% ±3.2** |

Table 4. Impact of limiting $|\mathcal{B}|$ on the mIoU results for the target data set ssTEM.

| Method | 1-shot ($\lvert\mathcal{B}\rvert = 100$) | 3-shot ($\lvert\mathcal{B}\rvert = 300$) | 5-shot ($\lvert\mathcal{B}\rvert = 500$) | 7-shot ($\lvert\mathcal{B}\rvert = 700$) | 10-shot ($\lvert\mathcal{B}\rvert = 1000$) |
|---|---|---|---|---|---|
| Entropy | 24.8%±1.8 | 40.8% ±0.5 | 45.3% ±1.3 | 46.7% ±1.4 | 48.7% ±1.3 |
| MC-dropout | 25.1% ±1.1 | 41.4% ±1.4 | 45.3% ±1.3 | 46.9% ±1.4 | 48.3% ±1.3 |
| Random | 46.2% ±2.2 | 55.5% ±3.4 | 61.3% ±3.5 | 63.3% ±4.2 | 64.4% ±2.3 |
| ***Ours*** | **48.5% ±1.8** | **58.7% ±3.4** | **63.6% ±2.6** | **65.3% ±2.2** | **68.4% ±4.2** |

Table 5. Impact of limiting $|\mathcal{B}|$ on the mIoU results for the target data set EM.

| Target: TNBC | | | | | |
|---|---|---|---|---|---|
| Approach | 1-shot | 3-shot | 5-shot | 7-shot | 10-shot |
| Expert | 32.4% | 41.0% | 40.0% | 41.1% | 39.6% |
| ***Ours*** | 50.3% | 50.8% | 49.0% | 51.4% | 53.0% |
| Target: ssTEM | | | | | |
| Approach | 1-shot | 3-shot | 5-shot | 7-shot | 10-shot |
| Expert | 38.0% | 55.3% | 66.0% | 68.9% | 72.4% |
| ***Ours*** | 52.6% | 63.2% | 60.5% | 67.3% | 70.2% |

Table 6. Effect on mIoU results of the expert selection and our approach on the target data set TNBC and ssTEM. We report the result for only one experiment i.e. we do not average over ten experiments.

| Target: B39 | | | | | |
|---|---|---|---|---|---|
| Method | 1-shot ($\|\mathcal{B}\| = 100$) | 3-shot ($\|\mathcal{B}\| = 300$) | 5-shot ($\|\mathcal{B}\| = 500$) | 7-shot ($\|\mathcal{B}\| = 700$) | 10-shot ($\|\mathcal{B}\| = 1000$) |
| Entropy | 80.9%±5.6 | 85.9% ±3.5 | 87.3% ±5.0 | 89.4% ±2.7 | 90.1% ±1.7 |
| MC-dropout | 72.1% ±3.0 | 71.3% ±3.2 | 73.1% ±3.3 | 76.6% ±2.6 | 77.9% ±3.9 |
| Random | **93.2% ±0.3** | **92.7% ±0.9** | **92.9% ±0.5** | 92.9% ±0.4 | 92.8% ±0.5 |
| ***Ours*** | 90.1% ±1.6 | 91.9% ±1.6 | 92.7% ±0.7 | **93.0% ±0.3** | **92.9% ±0.4** |
| Target: B5 | | | | | |
| Method | 1-shot ($\|\mathcal{B}\| = 100$) | 3-shot ($\|\mathcal{B}\| = 300$) | 5-shot ($\|\mathcal{B}\| = 500$) | 7-shot ($\|\mathcal{B}\| = 700$) | 10-shot ($\|\mathcal{B}\| = 1000$) |
| Entropy | 99.1%±0.1 | 99.1% ±0.1 | 99.1% ±0.1 | 99.1% ±0.1 | 99.0% ±0.5 |
| MC-dropout | 99.1% ±0.1 | 99.1% ±0.1 | 99.0% ±0.2 | 99.0% ±0.1 | 99.0% ±0.2 |
| Random | **99.1% ±0.0** | **99.1% ±0.0** | **99.0% ±0.2** | **99.1% ±0.0** | **99.1% ±0.0** |
| ***Ours*** | 98.9% ±0.0 | 99.0% ±0.0 | 99.0% ±0.0 | 99.0% ±0.0 | 99.0% ±0.0 |

Table 7. mIoU results for the target testing sets of B39, and B5. Best results are highlighted.