Full or Weak annotations?
An adaptive strategy for budget-constrained annotation campaigns
Supplementary Material

Javier Gamazo Tejero[1], Martin S. Zinkernagel[2], Sebastian Wolf[2]
Raphael Sznitman[1], Pablo Márquez Neila[1]

[1]University of Bern, [2]Inselspital Bern

{javier.gamazo-tejero, raphael.sznitman, pablo.marquez}@unibe.ch

{martin.zinkernagel, sebastian.wolf}@insel.ch

1. Implementation details

Weakly-supervised segmentation model hyperparameters: We used the U-Net segmentation model [5] for OCT, and the DeepLabV3 model [2] with a ResNet50 backbone on the SUIM, PASCAL, and Cityscapes datasets. For the U-Net, a classification head with max-avg pooling and one fully connected layer was appended at the end of the encoding module for the classification task. For the DeepLab-like models, we trained the entire ResNet-50 backbone on the classification task and then added the ASPP head for segmentation. In all cases, we used the cross-entropy loss for classification and the average of the Dice and cross-entropy losses for segmentation. Tab. 1 contains the details of batch sizes and optimizers.

Dataset	Model	Optimizer	Batch size
OCT	U-Net	Adam	8
VOC	DeepLabV3	SGD	16
SUIM	DeepLabV3	SGD	8
Cityscapes	DeepLabV3	SGD	16

Table 1. Hyperparameters and conditions for all experiments.

Algorithm hyperparameters: We measured the costs of annotations in terms of class-label equivalents setting $\alpha_c = 1$ and leaving only α_s as a hyperparameter of our method. We set to $\alpha_s = 12$ for all datasets following previous studies on crowdsourced annotations [1]. We fixed the number of iterative steps to $T = 8$ and the learning rate of the GP to 0.1. We set both the initial number of class annotations C_0 and segmentation annotations S_0 to 8% of the available labels for SUIM and Cityscapes. We reduced C_0 in OCT and S_0 in VOC to account for the higher number of labels available in those datasets, as detailed in Tab. 2.

Dataset	$C_0(\%)$	$S_0(\%)$	B_0
OCT	4	8	1'774
VOC	8	6	8'076
SUIM	8	8	1'586
Cityscapes	8	8	1'774

Table 2. Initial conditions for our method. B_0 calculated with $\alpha_s = 12$ and $\alpha_c = 1$.

2. Sensitivity to α_s

Fig. 1 shows aditional experiments on the sensitivity of our method to α_s in the considered datasets.

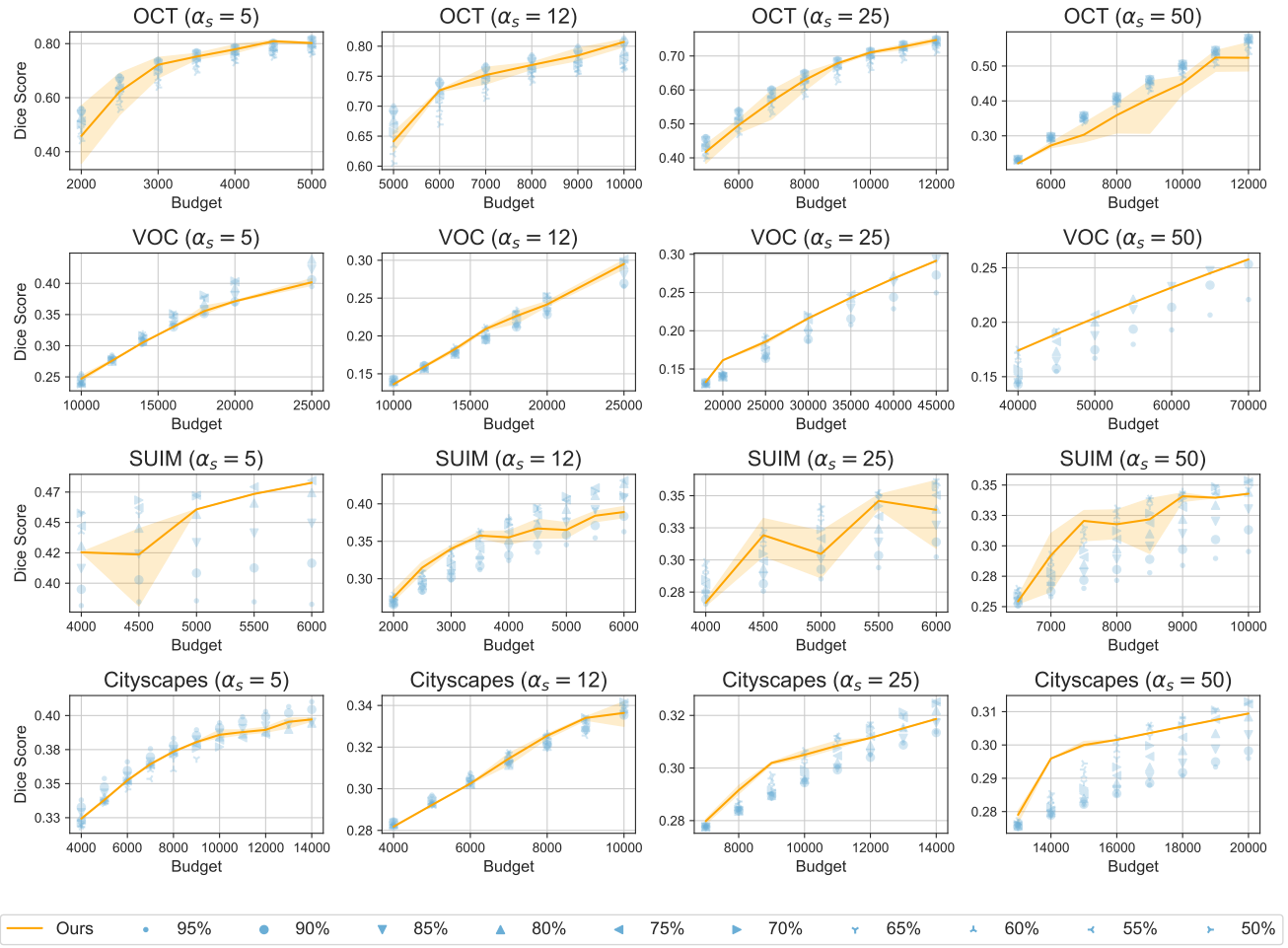

Figure 1. Performance our method with $\alpha_s = \{5, 12, 25, 50\}$ on four datasets. One-sigma error bars were computed from three seeds. Blue marks show the performance of fixed strategies, with labels indicating the percentage of the budget allocated to segmentation annotations.

3. Average performance

We report in Table 3 the average relative performance for each dataset and segmentation split. Our method is best on average and for two datasets (SUIM and VOC), while it performs on par with the best baseline for Cityscapes and OCT.

Seg. Split	95	90	85	80	75	70	65	60	55	50	Ours
Cityscapes	**73.9**	73.2	73.0	73.1	73.3	73.5	73.5	72.5	72.3	71.4	73.4
OCT	**95.4**	94.6	93.8	93.0	92.1	91.3	90.5	89.7	88.8	87.7	92.6
SUIM	72.1	74.1	76.7	79.2	80.9	81.3	79.7	78.1	76.7	75.1	**81.9**
VOC	44.3	44.0	44.3	44.9	45.6	46.0	43.1	43.3	40.9	40.8	**46.0**
Average	71.4	71.5	71.9	72.5	73.0	73.0	71.7	70.9	69.7	68.8	**73.5**

Table 3. Relative performance (in %) against full supervision for each dataset and segmentation split.

4. Surfaces

To explain the drop in performance for SUIM after a budget of 3'500, we computed the true surfaces for all datasets (Fig. 2). We observed that segmentation performance grows logarithmically for OCT, VOC, and Cityscapes, but not for SUIM after a certain number of class annotations. Since our GP assumes a logarithmic relation between dataset size and performance, this observation is particularly relevant to explain the decline in performance for SUIM. Most notably, and confirming our hypothesis, this decline is not seen with other values of α_s for this same dataset (Fig. 1). Due to the limited dataset size, larger values of α_s constrain the area of the surface that can be reached in our experiments. In the case of SUIM, $\alpha_s = 50$ translates to exploring zones where performance grows logarithmically with dataset size (low data regime).

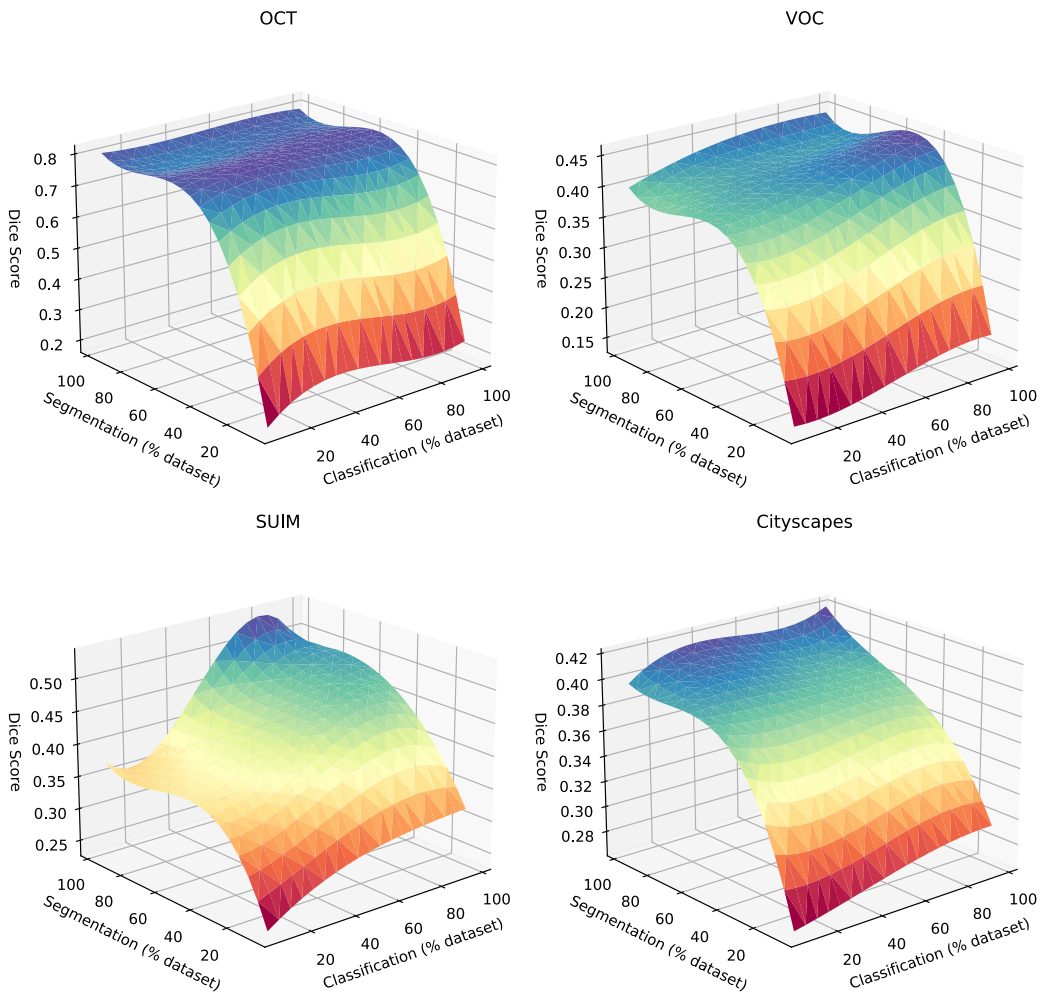

Figure 2. Segmentation performance grows logarithmically with training set size on Cityscapes, OCT, and VOC. This trend is not observed in the SUIM dataset.

5. Ground-truth approximation

To address the high computational demands of our experiments, we followed the procedure of [4] for ground-truth approximation. In particular, we built subsets of the training dataset by randomly sampling different proportions of the available annotated samples [3]. The proportions ρ_s and ρ_c for segmentation and classification samples, respectively, were chosen from the set $\{2\%, 4\%, 6\%, 8\%, 10\%, 20\%, 30\%, 40\%, 60\%, 80\%, 100\%\}$, for a total of $11 \times 11 = 121$ possible training subsets. For each subset, we trained the weakly-supervised segmentation model and measured its Dice score on a fixed segmentation test set. We finally interpolated these scores with third-order splines to obtain a surface of ground-truth Dice scores. This

procedure allowed efficient estimations of the Dice Score values without retraining a new model for each strategy.

The proportions ρ_s and ρ_c are relative to the total amounts of available annotated samples in each dataset, which are shown in Tab. 4.

Dataset	Segmentation	Classification
OCT	902	22'723
VOC	10'582	5'717
SUIM	1'525	1'525
Cityscapes	2'975	2'975

Table 4. Number of training images for each dataset and modality.

References

[1] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 549–565, Cham, 2016. Springer International Publishing. 1

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1

[3] Rafid Mahmood, James Lucas, David Acuna, Daiqing Li, Jonah Philion, Jose M Alvarez, Zhiding Yu, Sanja Fidler, and Marc T Law. How much more data do i need? estimating requirements for downstream tasks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 275–284, 2022. 3

[4] Rafid Mahmood, James Lucas, Jose M. Alvarez, Sanja Fidler, and Marc T. Law. Optimizing data collection for machine learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 10 2022. 3

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1