# Decomposition Sampling for Efficient Region Annotations in Active Learning: Supplementary Materials

Jingna Qiu[1,2] Frauke Wilm[1,3] Mathias Öttl[1,3] Jonas Utz[1]
Maja Schlereth[1] Moritz Schillinger[1] Marc Aubreville[4] Katharina Breininger[2]
[1]Friedrich-Alexander-Universität Erlangen-Nürnberg [2]Julius-Maximilians-Universität Würzburg
[3]MIRA Vision Microscopy GmbH [4]Hochschule Flensburg
jingna.qiu@fau.de katharina.breininger@uni-wuerzburg.de

## 1. Datasets and Tasks

### 1.1. BRACS

**BRACS** [4] contains 320 H&E-stained WSIs (train/val/test: $193/68/59$). Annotating a WSI requires complete RoI identification and classification. The full-annotation includes $3,163/602/626$ RoIs across the splits, with 1–119 RoIs per slide (median: 8). Notably, BRACS includes precancerous lesions (atypical ductal hyperplasia (ADH) and flat epithelial atypia (FEA)), which are clinically significant due to their progression risk [7]. The classification task is to categorize a RoI into 7 classes (normal, benign, usual ductal hyperplasia (UDH), ADH, FEA, ductal carcinoma in situ (DCIS), and invasive carcinoma). Additionally, a 3-class task is performed to distinguish higher-level categories: benign (normal, benign, UDH), atypical tumor (ADH, FEA), and malignant tumor (DCIS, invasive). 7-class results are provided in supplementary materials.

Annotation for the BRACS dataset followed a multi-step process [4]. Three pathologists first determined the most aggressive tumor subtype in each WSI as the image label. Then, each pathologist annotated a subset of WSIs, identifying and classifying RoIs. While exhaustive identification of all regions was not required, especially for certain classes like normal tissue, efforts were made to maintain balanced class distributions. Consequently, the number of RoIs per WSI varies from 1 to 119 (median: 8), with RoIs of varying sizes to encapsulate entire diagnostic lesions. This annotation process further demonstrate the practical meaning of image selection that we proposed: minimizing the number of WSIs requiring expert review streamlines the real-world annotation workflow and increases opportunities for obtaining multi-annotator assessment.

### 1.2. Cityscapes

The Cityscapes dataset [6] provides pixel-level urban scene segmentation for $2,975$ training and $500$ validation images across 19 classes (road, sidewalk, building, wall, fence, pole, traffic light, traffic sign, vegetation, terrain, sky, person, rider, car, truck, bus, train, motorcycle and bicycle). The dataset exhibits extreme class imbalance in full-annotation: the minority class (motorcycle) accounts for only $0.27\,\%$ of the annotated pixels of the majority class (road). The task involves 2-D segmentation across the 19 classes.

### 1.3. KiTS23

The KiTS23 dataset [11] includes 489 3-D abdominal CT images with manual annotations for kidneys, renal tumors, and renal cysts. The task involves hierarchical segmentation of the kidney region, the combined tumor and cyst regions, and the tumor alone. Accurate kidney tumor segmentation provides quantitative representations for risk stratification and treatment planning. Following [13], cases with confirmed or suspected faulty annotations (IDs: $23, 68, 125, 133, 15, 37$) were excluded. The CT scans contain 60–610 slices (median: 177). Dataset annotation involved identifying kidney regions in 3-D and delineating regions on axial planes [11].

## 2. full-annotation Benchmarks, Implementation Details, and Evaluation Metrics

### 2.1. RoI Classification on BRACS

We used HACT-Net [21] developed by the data provider, which constructs multi-level structural representations through cell- and tissue-graphs for breast cancer RoI subtyping. Nodes in the cell-graph represent nuclei, while tissue-graph nodes represent superpixel-based regions. For each node, features capturing morphological and spatial information are extracted to define inter-node interactions. The cell-graph and tissue-graph are processed independently by two distinct graph neural networks (GNNs). The cell-GNN processes a cell-node's feature and generates an embedding by aggregating information from neighboring nodes. The tissue-GNN creates a tissue-node embedding by operating on its feature and the embeddings of spatially located cell-nodes within it. The final representation for the RoI, obtained by summing all tissue-node embeddings, is passed through a multi-layer perceptron (MLP) followed by a softmax layer to classify the RoI into a specific breast cancer subtype. The full-annotation performance for the 7-class and 3-class tasks was benchmarked on the test set as weighted F1 scores of $0.5540$ and $0.7320$, respectively (mean of five runs). AL was used to annotate both training and validation set.

### 2.2. 2-D Segmentation on Cityscapes

We used InternImage [29] with UperNet [31] as the decoder for benchmarking, owing to its top leaderboard performance[2]. InternImage is an efficient convolutional neural network (CNN)-based foundation model featuring deformable convolutions (DCNs), which adaptively explore short- and long-range dependencies through learnable offsets and spatial aggregation scalars. It further splits each spatial aggregation process into multiple groups to learn diverse aggregation patterns for enhanced representation learning. We used InternImage-T for computational efficiency, given the iterative model training required in AL procedures. mIoU served as the evaluation metric. We trained maximally 64k iterations and stopped early when the validation performance stops improving for five consecutive 1k-iterations. The full-annotation performance achieved was $0.8096$ on the validation set (mean of five runs). AL was used to annotate the training set.

### 2.3. 3-D Segmentation on KiTS23

We used 3-D nnU-Net [14] for benchmarking the full-annotation performance on KiTS23, due to its dominant usage among top KiTS21 leaderboard entries [10]. Class-averaged DSC score was used as the evaluation metric. We attained full-annotation performance of $0.8263, 0.8762, 0.8516$ for 2-D, 3-D-lowres and 3-D-fullres configurations, respectively (mean of five-fold cross-validation). We therefore used the 3-D-lowres configuration for all following experiments. Note that nnU-Net contains a preprocessing step of detecting foreground voxels using the full-annotation for dataset intensity normalization. Since the full-annotation is not available in AL, we detected the foreground voxels with intensity values in the range of $[-200, 500]$ HU, resulting in a comparable DSC score of $0.8751$. Training on the fully annotated dataset with $1000$ epochs, as in [14], takes approximately seven hours using one NVIDIA A100 Tensor Core GPU. To improve computational efficiency, we reduced the training epochs to align with the total number of annotated regions in AL experiments, up to a maximum of $1000$ epochs. AL was used to annotate the training set. AL was used to annotate the training set for each corresponding cross-validation fold, and the average performance across the five folds was reported.

## 3. Comparison Methods

Here we provide more implementation details of comparison methods.

**RAND** : It randomly selects $n_{\text{image}}$ images and then choose $n_{\text{region}}$ non-overlapping regions at random locations within each image. Of note, foreground detection was performed before region selection on KiTS23.

[2] https://www.cityscapes-dataset.com/benchmarks/

**UNCERT**: We compare to classic *entropy*-based method for RoI classification and 2-D segmentation, where model outputs are processed with softmax, and to *least confidence*-based method for the 3-D segmentation task with hierarchical classes, where model outputs are processed with sigmoid. These methods first select images with the highest average predictive uncertainty across all pixels (segmentation) or RoIs (RoI classification), then choose the most uncertain regions within those images. For segmentation tasks, this involves splitting the image into overlapping regions with a stride of one pixel, calculating the average uncertainty of pixels within each region as the region uncertainty, and using non-maximum suppression to select $n_{\text{region}}$ non-overlapping regions with the highest uncertainties. For RoI classification, the region uncertainty is computed directly from its prediction.

**DIVERS**: First, the $n_{\text{image}}$ most uncertain images are first selected. Then, $3 \times n_{\text{region}}$ regions with the highest uncertainties are identified from each image to form a pool, from which $n_{\text{image}} * n_{\text{region}}$ regions are ultimately chosen using clustering (**DIVERS(cluster)**) or core-set identification (**DIVERS(core-set)**) based on region features. This approach helps to avoid selecting similar regions across different images. Features from the bottleneck layer are used in segmentation tasks, while features from the penultimate layer before the classifier are used in RoI classification tasks. For the clustering-based method, we perform k-means clustering with the number of clusters set to $n_{\text{image}} * n_{\text{region}}$, and select the region with the highest uncertainty in each cluster, following [16]. For the core-set-based method, we followed [32] to incrementally expand a set that maximally covers the latent space by iteratively including the region with the largest minimal distance to the already selected regions.

**BADGE** [1] selects informative samples by constructing gradient embeddings for each region, and applying clustering to promote both uncertainty and diversity. For RoI classification task, we follow the official implementation by weighting the feature of the penultimate layer by the discrepancy between predicted probabilities and one-hot pseudo-labels. For segmentation task, we replace with the feature obtained from the bottleneck layer, perform max pooling, and weight it with the averaged prediction discrepancy across all pixels in the region.

## 4. Additional Results on BRACS

Due to page limits, we report only the 3-class task in the main text, with 7-class results in Fig. 10 in the supplementary materials. The 7-class task is harder because of strong ambiguities among neighboring classes with shared morphology [4], resulting in lower full-annotation performance. Still, DECOMP maintains a clear sampling efficiency advantage over UNCERT and RAND.

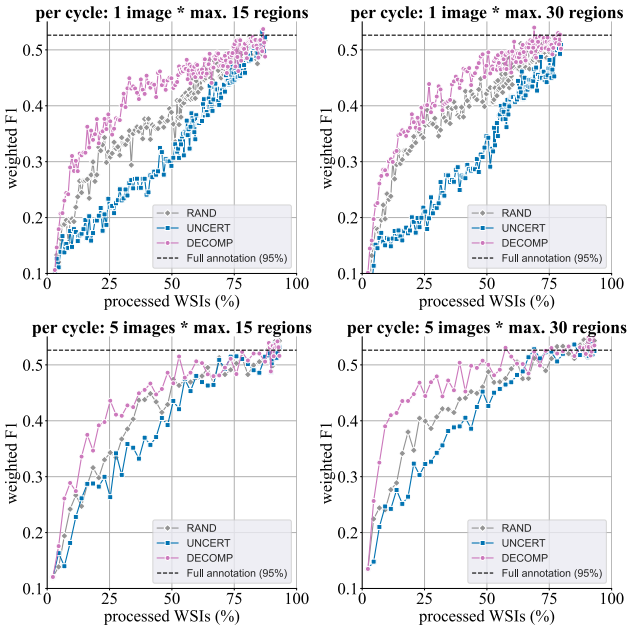


Figure 10. Results on the BRACS dataset for the 7-class task across $160, 140, 50, 50$ cycles for AL hyperparameter combinations of $n_{\text{image}} \in \{1, 1, 5, 5\}$ and $n_{\text{region}} \in \{15, 15, 30, 30\}$, respectively. Weighted F1 as a function of annotated WSIs (%) for different sampling methods. Mean over five runs.

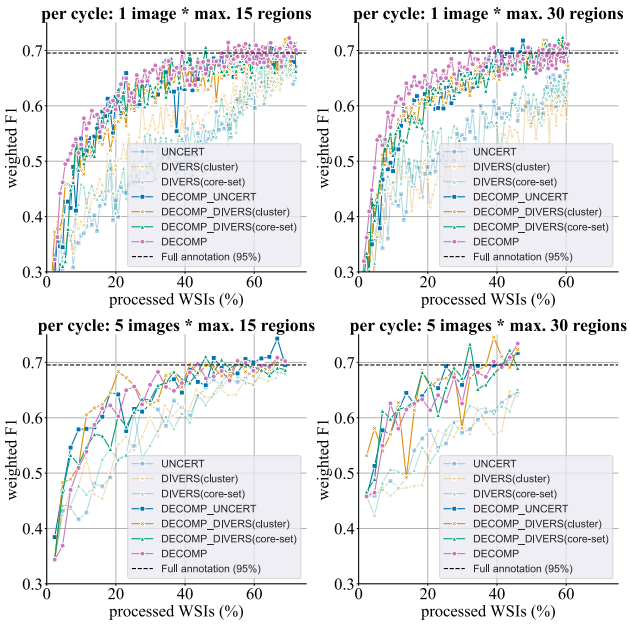


Figure 11. Effect of image selection in DECOMP on the BRACS dataset.

Figure 11 complements the Cityscapes ablation by showing the effect of DECOMP's image selection on BRACS. Similar to Cityscapes, switching image selection in UNCERT, DIVERS(cluster), and DIVERS(core-set) to DECOMP yields large gains, across all annotation budget settings.