This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

AcZeroTS: Active Learning for Zero-shot Tissue Segmentation in Pathology Images

Jiao Tang Junjie Zhou Bo Qian Peng Wan Yingli Zuo Wei Shao* Daoqiang Zhang
The College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics
The Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education
tangjiao@nuaa.edu.cn, shaowei20022005@nuaa.edu.cn, dqzhang@nuaa.edu.cn

Abstract

*Tissue segmentation in pathology images is crucial for computer-aided diagnostics of human cancers. Traditional tissue segmentation models rely heavily on large-scale labeled datasets, where every tissue type must be annotated by experts. However, due to the complexity of tumor microenvironment, collecting annotations for all possible tissue types is challenging, which makes the traditional methods ineffective in segmenting unseen tissue types with zero training samples. With the rapid development of vision-language models (VLMs), recent studies extend their powerful zero-shot capabilities to pixel-level segmentation tasks, where the model is trained only on seen classes but can perform tissue segmentation on both seen and unseen categories in the testing phase. However, these VLM-based zero-shot segmentation models still require substantial annotation efforts on seen classes. To attach desirable segmentation performance on both seen and unseen categories with limited labeled data, we propose **AcZeroTS**, a novel active learning framework for zero-shot tissue segmentation in pathology images. Specifically, AcZeroTS is built on a VLM-based prototype-guided zero-shot segmentation model called **ProZS**. We introduce a novel active selection criterion to select the most valuable samples for annotation on seen classes, which not only considers both uncertainty and diversity of unlabeled samples, but also ensures that the generated prototypes of ProZS can effectively summarize both seen and unseen classes during inference. We evaluate our method on two pathology datasets (TNBC and HPBC) as well as a natural dataset (Pascal VOC 2012), and the experimental results demonstrate the superiority of our method in comparison with the existing studies.*

1. Introduction

Histological examination of tissue pathology slides remains the gold standard for cancer diagnosis and prognosis [42]. However, visual examinations performed by pathologists are typically subjective, resulting in variability across different observers. To ensure objective and consistent evaluations, computerized quantitative image analysis is highly desirable [43]. Among different pathological image analysis tasks, tissue segmentation aims at identifying diverse tumor micro-environment (TME) components that are essential for understanding cancer progression [44].

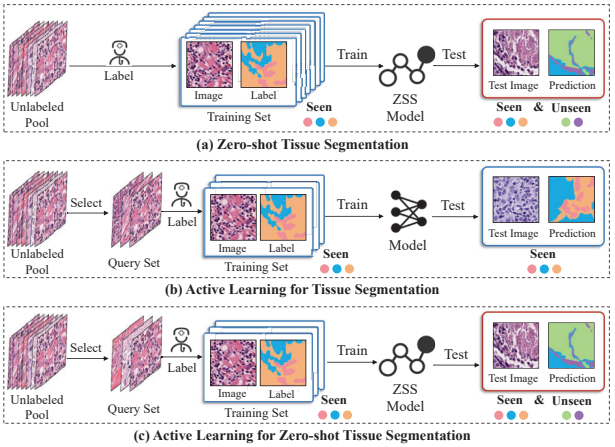


Figure 1. (a) Traditional zero-shot tissue segmentation models require labeling large amounts of seen-class tissues. (b) Traditional active learning methods are ineffective for tissue segmentation on unseen classes. (c) We design an active learning framework for zero-shot tissue segmentation that only selects the most valuable samples of seen classes for annotation but still achieves desirable segmentation performance on both seen and unseen tissue types.

Over the past decades, deep learning technology has achieved remarkable success in computer vision, which also boosts considerable interests in designing deep tissue segmentation models [15, 37]. Current tissue segmentation models are mostly trained in a supervised manner that requires heavy annotation efforts from the pathologists. However, due to the complex structure of the TME components revealed in pathology images, it is challenging to collect annotations for all possible tissue types. For instance, most of the tissue segmentation models focus on distinguishing tumor and stromal regions in pathology images given the la-

[1]Corresponding Author

beled data. For the purpose of providing more insights into tumor progression and metastasis, we also need to analyze the spatial organization of lymphocytes that are hard for annotation but often considered as an important marker of the immune system to cancer [48]. However, directly adopting the tumor/stromal segmentation models for the identification of lymphocyte regions will yield unsatisfied results.

To address the above issue, several zero-shot segmentation (ZSS) methods are proposed and most of them rely on the Segment Anything Model (SAM) [19, 24, 32]. Although SAM's capability for tissue segmentation is appealing, it initially segments every parts and still requires additional prompt information to clarify the target unseen classes. Inspired by the fast development of vision-language models (VLMs), recent methods transfer their powerful zero-shot generalization capability from the image-level classification tasks to the pixel-level segmentation tasks, where the model is trained only on seen classes and can segment unseen classes that are not available during training. For instance, the studies in [20, 47] introduce the text information and extend the CLIP [29] model for ZSS tasks. In addition, Aleem *et al.* [1] integrate SAM and CLIP into a unified framework for ZSS on different medical image datasets.

However, training an accurate and reliable zero-shot tissue segmentation model still requires amounts of high-quality labeled data on seen classes, leading to high annotation costs for pathologists [34] (Fig. 1 (a)). In order to reduce the labeling costs, several active learning methods have been proposed [18, 40, 41] that aim at querying the most valuable samples from the unlabeled pool for annotation. However, these research efforts have not investigated how to design effective query strategies in the ZSS tasks (Fig. 1 (b)). Intuitively, in the setting of ZSS, the query strategy of active learning should attach equal significance to both seen and unseen categories with limited labeled data (Fig. 1 (c)).

Based on the above considerations, in this work, we propose **AcZeroTS**, a novel active learning framework for zero-shot tissue segmentation in pathology images, which is built on a VLM-based prototype-guided zero-shot segmentation model called **ProZS**. AcZeroTS can select the most valuable samples of seen classes for annotation and improve the performance on both seen and unseen classes during the inference stage. The main contributions of this study are listed as follows:

(1) To the best of our knowledge, this is the first active learning framework designed for zero-shot segmentation, aimed at optimizing sample selection to improve model generalization across both seen and unseen classes with limited labeled data.

(2) We design three criteria and combine them together for query selection. Specifically, the criteria of *Evidence-based Uncertainty of the Unlabeled Data (EUN)* and *Contrast Among Labeled and Unlabeled Data (COT)* consider both uncertainty and diversity of unlabeled samples. More importantly, the criterion of *Complete Summarization Ability of the Prototypes (CSA)* measures the summarization ability of the generated prototypes on both seen and unseen tissue types.

(3) We evaluate the effectiveness of our method on two pathology datasets (TNBC and HPBC) as well as a natural dataset (PASCAL VOC 2012), and the experimental results demonstrate that AcZeroTS can achieve higher segmentation accuracy on both seen and unseen categories with lower annotation burden in comparison with the existing studies.

2. Related Work

Zero-shot Segmentation (ZSS) aims to accurately segment unseen classes with zero training samples. Previous works like ZS3 [5], SPNet [39], CaGNet [10], Joint [3], SIGN [7] and [13, 45] have tried to tackle the task by enhancing the generalization ability of semantic mapping from seen to unseen classes. With the development of vision foundation model Segment Anything Model (SAM) [19], the studies in [24, 32] adopt SAM to realize ZSS, but most of them ignore the fact that the semantic information is also helpful for the ZSS task. Inspired by the rapid progress of vision-language models (VLMs), recent methods extend their strong zero-shot generalization ability from image-level classification to pixel-level segmentation. For instance, Aleem *et al.* [1] incorporate SAM and CLIP into a unified framework for zero-shot segmentation on medical images. ZegFormer [8], ZegCLIP [47], and TagCLIP [20] further extend the CLIP model and introduce text information for ZSS tasks. Moreover, MaskCLIP+ [46] applies CLIP to generate pseudo annotations on novel classes for self-training. However, it is still challenging to transfer the zero-shot ability of the vision-language foundation models from image-level to pixel-level, and the requirement for large amounts of labeled data on seen classes will bring heavy annotation burden for experts.

Active Learning for Semantic Segmentation aims at reducing labeling costs by querying the most valuable samples for pixel-level annotation. In general, the query strategy can be divided into two aspects, i.e., uncertainty and representativeness based criteria. The uncertainty-based criterion is a widely used strategy in active learning, which selects the most informative samples by measuring the model's uncertainty about its predictions [12, 33, 41]. For instance, Xie *et al.* [40] present an active learning algorithm to deal with performance limitations of domain adaptive semantic segmentation at minimal labeling costs. Cai *et al.* [6] develop a super-pixel-based approach for active learning in semantic segmentation that takes into account more realistic click-based annotation costs. On the other hand,

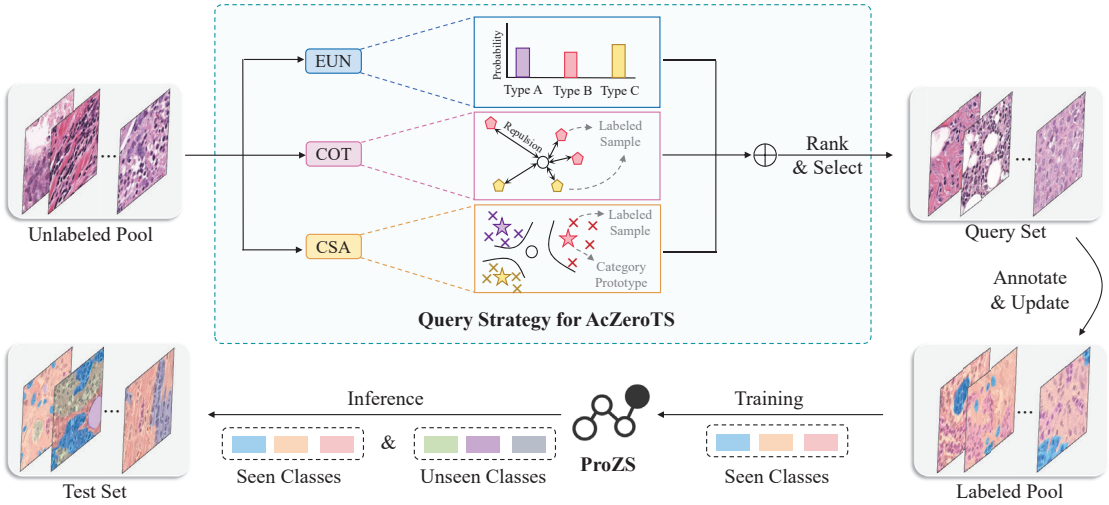


Figure 2. Flowchart of the proposed AcZeroTS, which is built upon a well designed prototype-guided ZSS model, i.e., ProZS (shown in Fig. 3). The query strategy for AcZeroTS consists of three criteria in the form of Evidence-based Uncertainty of Unlabeled Data (EUN), Contrast Among the Labeled and Unlabeled Data (COT) and Complete Summarization Ability of the Prototypes (CSA). In each cycle of active learning for AcZeroTS, we query the most valuable samples from the seen tissue categories for annotation. Based on the queried data, we then update ProZS to improve the segmentation performance on both seen and unseen categories.

the representativeness-based query strategy in active learning aims to select samples that best represent the overall data distribution [31]. Another study called MADA [27] adopts multiple anchors instead of a single centroid to select more representative samples for annotation. However, these active learning methods are not designed for segmenting novel categories that have not been seen in the training set. The active learning framework designed for zero-shot segmentation that can improve model generalization ability on both seen and unseen classes with limited labeled data has not been explored.

3. Method

3.1. Problem Formulation

We follow the inductive setting of generalized zero-shot segmentation (GZSS) [39] which requires to segment both seen and unseen classes $C^s \cup C^u$ after training on the seen classes C^s, $C^s \cap C^u = \varnothing$. Suppose that we have a small labeled pool $\mathcal{L} = \{\boldsymbol{x}_i^L\}_{i=1}^{n^L}$ with n^L samples and a large unlabeled pool $\mathcal{U} = \{\boldsymbol{x}_j^U\}_{j=1}^{n^U}$ with n^U samples ($|\mathcal{L}| \ll |\mathcal{U}|$), where the pixels in $\mathcal{L}$ and $\mathcal{U}$ belong to seen categories in the training process. The goal of our framework aims at querying the most valuable samples from $\mathcal{U}$ for annotation, followed by adding them into $\mathcal{L}$ for ZSS model re-training, which can enhance the segmentation performance on both seen and unseen classes during inference while minimizing annotation costs.

3.2. Architecture Overview

We show the flowchart of the proposed active learning algorithm AcZeroTS in Fig. 2. AcZeroTS is built upon a prototype-guided ZSS model ProZS (shown in Fig. 3). At each iteration of active learning for AcZeroTS, we select the most valuable samples from the seen tissue categories for annotation by considering the following three criteria: Evidence-based Uncertainty of Unlabeled Data (EUN), Contrast Among the Labeled and Unlabeled Data (COT) and Complete Summarization Ability of the Prototypes (CSA), where EUN and COT consider both uncertainty and diversity of unlabeled samples, and the criterion of CSA measures the summarization ability of the generated prototypes on both seen and unseen tissue categories. By incorporating the queried data for updating ProZS, we could achieve better segmentation results on both seen and unseen tissue categories in the test set.

3.3. ProZS: Prototype-guided Zero-shot Segmentation Model

Fig. 3 presents the framework of ProZS, which integrates the pre-trained pathology visual-language model, i.e., PLIP [14] with prototype learning for zero-shot segmentation on pathology images. It comprises a vision pathway for extracting visual features and a text pathway for encoding semantic information to guide segmentation.

Vision Pathway: We split the input image into M patches with token embeddings $\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_M] \in \mathbb{R}^{M \times d_v}$, where d_v is the token dimension. Suppose that the image encoder is consisted of L layers, and the multi-head attention module in the l-th layer can be formulated as:

$$[\boldsymbol{g}^l, \boldsymbol{P}^l] = Layer^l([\boldsymbol{g}^{l-1}, \boldsymbol{P}^{l-1}]), \tag{1}$$

where $\boldsymbol{g}^l$ is the [CLS] token embedding and $\boldsymbol{P}^l$ =

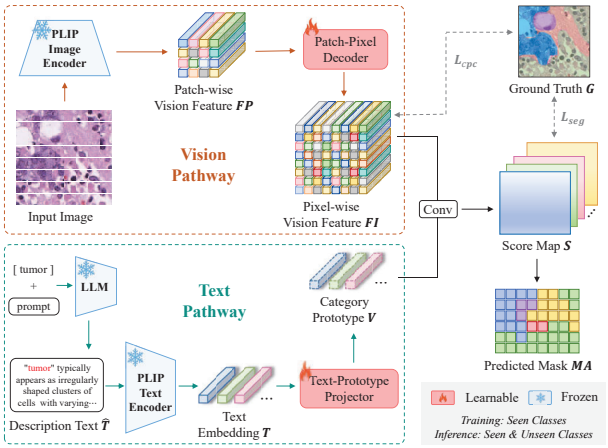


Figure 3. Framework of prototype-guided zero-shot segmentation model based on pathology vision-language model.

$[\boldsymbol{p}_1^l, \boldsymbol{p}_2^l, \cdots, \boldsymbol{p}_M^l]$ represents patch embeddings at layer l. We aggregate patch embeddings across layers as $\hat{\boldsymbol{P}} = \sum_{l=1}^{L} \boldsymbol{P}^l$ and reshape it into a patch-wise feature map $\boldsymbol{FP} \in \mathbb{R}^{H' \times W' \times d_v}$, where $H' = W' = \sqrt{M}$. To achieve pixel-wise segmentation, we design a patch-to-pixel decoder that extends $\boldsymbol{FP}$ into a pixel-wise feature map $\boldsymbol{FI} \in \mathbb{R}^{H \times W \times d}$ with matched resolution to the original image. Here, d represents the dimension of pixel-level features. The structure of the patch-to-pixel decoder is detailed in Fig. 4 (a), including convolution and up-sampling blocks that can effectively reconstruct the pixel-wise feature.

Text Pathway: Instead of generic text prompts like "a photo of {}", we generate description texts for different tissue categories by considering their appearance attributes using a large language model (Details in *Supplementary Materials* (Sec. 6)). Given the description set $\hat{\boldsymbol{T}} = \{\hat{\boldsymbol{t}}_1, \hat{\boldsymbol{t}}_2, \cdots, \hat{\boldsymbol{t}}_{N_c}\}$, where N_c is the number of categories ($N_c = |C^s|$ during training and $N_c = |C^s| + |C^u|$ during inference), we use the text encoder of PLIP to obtain the semantic embeddings of text descriptions $\boldsymbol{T} = [\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_{N_c}] \in \mathbb{R}^{N_c \times d_t}$ with the dimension of d_t. Then, we train a text-to-prototype projector in the form of three linear layers that can map the semantic representation $\boldsymbol{T}$ to the category prototypes $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_{N_c}] \in \mathbb{R}^{N_c \times d}$ in vision-semantic space, enabling flexible representation of unseen classes during inference.

Prototype-guided Segmentation in Vision-semantic Space: Since the pixel-wise feature $\boldsymbol{FI} \in \mathbb{R}^{H \times W \times d}$ and category prototypes $\boldsymbol{V} \in \mathbb{R}^{N_c \times d}$ involve vision and semantic information, we conduct the tissue segmentation task by the combination of them via prototype learning. Specifically, we treat $\boldsymbol{V}$ as convolutional kernels and then convolve it with $\boldsymbol{FI}$ .The softmax operator is adopted to generate a probability score map $\boldsymbol{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_{N_c}\} \in \mathbb{R}^{H \times W \times N_c}$ over semantic classes. We use focal loss [35] as the segmentation loss:

$$\mathcal{L}_{seg} = \mathcal{L}_{focal}(\boldsymbol{S}, \boldsymbol{G}), \tag{2}$$

where $\boldsymbol{G}$ is the ground truth mask. To enhance the discriminability of the pixel-level features, we introduce a category-wise pixel-level contrastive loss $\mathcal{L}_{cpc}$ to learn pixel-level representations across different categories. As shown in Fig. 4 (b), pixels are grouped into N_c sets, i.e., $\boldsymbol{SP} = \{\boldsymbol{sp}_1, \boldsymbol{sp}_2, \cdots, \boldsymbol{sp}_{N_c}\}$ based on ground truth labels, we denote the centroid representation of each group as:

$$\boldsymbol{cp}_i = \frac{1}{|\boldsymbol{sp}_i|} \sum_{t \in \boldsymbol{sp}_i} \boldsymbol{FI}^t, \tag{3}$$

where t indexes pixels. For the purpose of aligning the pixel-level representations with their corresponding centroids, we design a contrastive loss below:

$$\mathcal{L}_{cpc} = -\sum_{i=1}^{N_c} \sum_{t \in \boldsymbol{sp}_i} \frac{\alpha \cdot \boldsymbol{FI}^t \cdot \boldsymbol{cp}_i}{\sum_{j=1}^{N_c} \alpha \cdot \boldsymbol{FI}^t \cdot \boldsymbol{cp}_j}, \tag{4}$$

where α is the temperature factor. By the combination of Eq. (2) and Eq. (4), the overall loss for our ProZS can be formulated as:

$$\mathcal{L}_{overall} = \mathcal{L}_{seg} + \lambda \mathcal{L}_{cpc}, \tag{5}$$

where λ balances the contribution of $\mathcal{L}_{seg}$ and $\mathcal{L}_{cpc}$. In the testing stage, pixel-wise labels are assigned via:

$$\boldsymbol{MA}^t = \underset{i \in [1, N_c]}{\arg\max} \{\boldsymbol{s}_i^t\}. \tag{6}$$

3.4. Active Learning Selection Strategy for Zero-shot Segmentation

In this section, we introduce our active learning selection strategy, designed to optimize sample selection for seen classes while enhancing the generalization ability on both seen and unseen categories. It consists of three key criteria: 1) Evidence-based Uncertainty of Unlabeled Data (EUN), 2) Contrast Among the Labeled and Unlabeled Data (COT), and 3) Complete Summarization Ability of the Prototypes (CSA). EUN and COT focus on selecting diverse and uncertain samples, while CSA ensures that the generated prototypes in ProZS effectively summarize both seen and unseen categories during inference.

Evidence-based Uncertainty of the Unlabeled Data: Considering the image with rich information is crucial for the generalization ability of the model [22], we adopt an evidence-based uncertainty estimation technique introduced in [11] to avoid the issue that the traditional entropy-based methods tend to produce inflated confidence values even for incorrect predictions [26]. Specifically, for the unlabeled image $\boldsymbol{x}_j^U$, we calculate its score map $\hat{\boldsymbol{S}}_j$ according

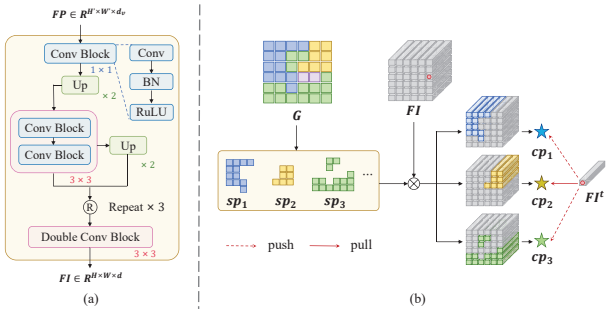


Figure 4. (a) The structure of the patch-to-pixel decoder. (b) Visualization of the category-wise pixel-level contrastive loss.

to the prototype learning introduced in Sec. 3.3 without applying softmax. Then we follow the study in [9] that adopts the ReLU activation function to derive its evidence map $\boldsymbol{E}_j = \{\boldsymbol{e}_{j,1}, \boldsymbol{e}_{j,2}, \cdots, \boldsymbol{e}_{j,N_c}\} \in \mathbb{R}^{H\times W\times N_c}$. Following the subjective logic in [17], we connect the pixel-wise evidence to the Dirichlet distribution parameters $\boldsymbol{\beta}_{j,k} = \boldsymbol{e}_{j,k}+1$ with strength $\boldsymbol{Q}_j = \sum_{k=1}^{N_c} \boldsymbol{\beta}_{j,k}$, by which we can assign a belief mass $\boldsymbol{q}_{j,k} = \boldsymbol{e}_{j,k}/\boldsymbol{Q}_j$ for each class k. In essence, the increased evidence supporting a class results in a greater belief mass being allocated to it. Based on it, the uncertainty map is formulated as $\boldsymbol{UM}_j = 1 - \sum_{k=1}^{N_c} \boldsymbol{q}_{j,k}$, and thus the *uncertainty score* for $\boldsymbol{x}_j^U$ can be calculated by averaging the score of all its invovled pixels denoted as:

$$US_j = \frac{1}{H\times W} \sum_{t\in \boldsymbol{UM}_j} \boldsymbol{UM}_j^t. \tag{7}$$

Contrast among the Labeled and Unlabeled Data: To increase the discrepancy of data, we resort to certain contrast samples in the unlabeled pool that exhibit higher feature similarities with labeled images but result in significantly different category distribution. Intuitively, such samples are located near the decision boundary that are valuable for annotation. Specifically, for each learned pixel feature $\boldsymbol{FI}_j^t$ in the unlabeled image $\boldsymbol{x}_j^U$, we firstly calculate its distance to all labeled pixel features and identify its K nearest neighbors $\{\boldsymbol{FL}_{j,1}^t, \boldsymbol{FL}_{j,2}^t, \cdots, \boldsymbol{FL}_{j,K}^t\}$. Subsequently, we compute the Kullback–Leibler (KL) divergence between the category distribution $p(\boldsymbol{y}|\boldsymbol{FI}_j^t)$ and $p(\boldsymbol{y}|\boldsymbol{FL}_{j,n}^t)$, $n = 1, \cdots, K$. Then, the contrast map $\boldsymbol{CM}_j$ of $\boldsymbol{x}_j^U$ is formulated as:

$$\boldsymbol{CM}_j^t = \frac{1}{K}\sum_{n=1}^{K} \mathrm{KL}[p(\boldsymbol{y}|\boldsymbol{FI}_j^t)||p(\boldsymbol{y}|\boldsymbol{FL}_{j,n}^t)]. \tag{8}$$

According to Eq.(8), the *contrast score* of $\boldsymbol{x}_j^U$ can be calculated by averaging all its involved pixels on $\boldsymbol{CM}_j$ as:

$$CS_j = \frac{1}{H\times W} \sum_{t\in \boldsymbol{CM}_j} \boldsymbol{CM}_j^t. \tag{9}$$

Complete Summarization Ability of the Prototypes: To enhance the prototype summarization ability that can help identify unseen classes during inference, we prefer selecting samples that are distant from all category prototypes. Specifically, for the unlabeled image $\boldsymbol{x}_j^U$ with pixel feature map $\boldsymbol{FI}_j$, we calculate the distance between each pixel and N_c category prototypes in $\boldsymbol{V}$, producing a distance map $\boldsymbol{D}_j \in \mathbb{R}^{H\times W\times N_c}$, whose element $\boldsymbol{D}_{j,k}^t =< \boldsymbol{FI}_j^t, \boldsymbol{v}_k >$ measures the Euclidean distance between the t-th pixel of $\boldsymbol{FI}_j$ and the k-th category prototype. To identify samples far from all prototypes, we apply a minimum distance maximization approach, aiming to select the sample whose distance from its nearest prototype is father. Therefore, we calculate a minimum distance map $\boldsymbol{DM}_j \in \mathbb{R}^{H\times W}$ with $\boldsymbol{DM}_j^t = \min_{1\le k\le N_c} \boldsymbol{D}_{j,k}^t$. Then we compute the *prototype score* of $\boldsymbol{x}_j^U$ by averaging all its involved pixels on $\boldsymbol{DM}_j$ as:

$$PS_j = \frac{1}{H\times W} \sum_{t\in \boldsymbol{DM}_j} \boldsymbol{DM}_j^t. \tag{10}$$

Based on Eq. (10), we prioritize samples with higher prototype scores (i.e., those farther from all prototypes) to enhance the summarization ability of the prototypes, which will help generate more representative prototypes for both seen and unseen classes. Accordingly, these prototypes can improve the segmentation performance of unseen classes when convolved with the pixel-level feature map during inference.

Query Strategy: We combine uncertainty score US_j (Eq. (7)), contrast score CS_j (Eq. (9)), and prototype score PS_j (Eq. (10)) to jointly select samples for annotation. Intuitively, we aim for samples with higher uncertainty and diversity, while also enhancing the summarization ability of prototypes. Therefore, we apply the min-max normalization to scale scores uniformly, and the final *query score* QS_j is:

$$QS_j = US_j + PS_j + CS_j. \tag{11}$$

Based on the query score, we rank and select top N images from the unlabeled pool for annotation, which are then added to the labeled set for the next cycle of active learning.

4. Experiment

4.1. Dataset

To evaluate the effectiveness of our method, we conduct extensive experiments on two pathology datasets (TNBC and HPBC). We also extend our method on a natural dataset (Pascal VOC 2012) to further verify its effectiveness.

TNBC (Triple-Negative Breast Cancer) [2] is a public available dataset for tissue segmentation, which contains 8,783 pathological image patches with 6 categories: tumor (tum, 45%), stroma (str, 34%), lymphocytic infiltrate (lym,

Table 1. Comparison of AcZeroTS with previous active learning methods on TNBC. 'C' represents the active learning cycle.

Method	1 unseen						2 unseen						3 unseen					
	C1	C2	C3	C4	C5	C6	C1	C2	C3	C4	C5	C6	C1	C2	C3	C4	C5	C6
Random	9.5	14.4	32.4	35.9	41.6	46.0	4.8	10.2	25.1	30.1	31.9	32.8	16.8	19.8	24.0	32.4	33.8	34.9
Entropy [12]	20.3	38.9	40.0	44.7	45.7	46.6	10.4	19.1	25.9	32.0	33.5	34.6	17.0	26.8	28.5	33.9	34.3	36.5
Coreset [31]	22.1	42.7	44.6	45.7	45.5	46.7	13.2	33.6	32.8	34.9	36.4	37.0	21.1	27.9	32.2	34.8	35.2	36.3
DEAL [41]	21.3	40.9	43.5	47.5	46.1	47.9	14.3	29.1	31.3	33.0	36.9	37.2	20.6	27.5	31.1	33.9	34.5	36.7
MADA [27]	24.1	44.4	45.5	46.8	47.6	48.5	17.4	34.7	36.8	37.5	39.4	39.1	23.2	29.1	32.3	33.6	35.0	35.9
RIPU [40]	22.8	44.5	46.0	47.1	49.0	49.3	18.3	33.1	35.4	36.5	37.8	39.4	21.6	29.5	31.2	33.8	34.9	35.8
S4AL [30]	23.0	42.9	45.8	47.9	47.7	48.9	15.7	34.2	35.9	35.7	37.2	38.8	22.8	29.4	33.1	34.9	35.7	36.7
PBAL [28]	23.2	41.4	42.9	46.5	48.7	48.6	14.5	23.5	31.9	35.6	37.0	38.6	18.9	29.8	30.3	33.5	34.4	36.5
AcZeroTS (ours)	**24.3**	**48.3**	**52.3**	**53.3**	**53.6**	**54.1**	**18.5**	**38.5**	**39.8**	**41.8**	**41.0**	**41.8**	**32.6**	**35.3**	**37.7**	**37.5**	**38.0**	**38.9**

10%), necrosis or debris (nec, 7%), fat (2%), and plasma cells (pla, 2%). To construct the zero-shot learning (ZSL) environment, we set the number of unseen classes as **1 unseen:** [fat], **2 unseen:** [fat, pla], **3 unseen:** [fat, pla, nec].

HPBC (Her2-Positive Breast Cancer) is an in-house dataset including 4,956 patches with 4 categories: background (24%), stroma (str, 56%), lymphocytic (lym, 4%), and tumor (tum, 16%). We set [lym] as the **unseen class**.

Pascal VOC 2012 contains 10,582 augmented images for training and 1,449 images for validation with 20 categories. We follow the same unseen class setting in previous works [20, 47] that divide the dataset into 15 seen classes and 5 unseen classes.

4.2. Experimental Setting

Implementation Detail. For all datasets, we randomly allocate 70% of data for training and the rest for testing. In the active learning setup, 5% of the training set is randomly selected as the initial labeled pool, while the remaining samples form the unlabeled pool. In each active learning cycle, we query a batch of samples equivalent to 15% of the training set for annotation. For the ZSS model, all experiments on two pathology datasets use the pre-trained PLIP ViT-B/32 model with an input size of 224 $\times$ 224, producing a 7 $\times$ 7 feature map with 768 dimensions. We use GPT-3 [4] as the LLM to generate descriptions of different tissue categories. The patch-to-pixel decoder reduces the feature dimension from 768 to 64. Since text embeddings have a 512-dimensional representation, we use the text-to-prototype projector to map them to 64 dimensions. We empirically set the temperature factor α (Eq. (4)), the regularization parameter λ (Eq. (5)), and the number of neighbors K (Eq. (9)) as 0.5, 0.01, and 5, respectively. For Pascal VOC 2012, we replace PLIP encoders with CLIP's. All experiments use a batch size of 64, training for 100 epochs. We use SGD as the optimizer with a learning rate of 0.007 and weight decay of 5e-4.

Evaluation Metric. Following previous works [20, 47], we evaluate our model using mean of class-wise Intersection over Union (mIoU). In GZSS, seen-class mIoU is commonly much higher than unseen-class mIoU, dominating the overall mIoU [45]. Therefore, we follow [38] and use harmonic mIoU (hIoU) to evaluate segmentation performance, which balances seen and unseen classes and can better demonstrate overall performance of ZSS. Detailed definition are provided in *Supplementary Materials* (Sec. 7.1).

Table 2. Comparison of AcZeroTS with previous active learning methods on HPBC. 'C' represents the active learning cycle.

Method	C1	C2	C3	C4	C5	C6
Random	16.6	26.8	30.2	33.5	36.0	37.8
Entropy [12]	15.2	28.3	31.3	35.1	37.0	38.3
Coreset [31]	22.1	30.3	31.8	35.7	38.7	39.5
DEAL [41]	13.9	27.0	32.6	34.1	37.3	39.1
MADA [27]	20.1	28.5	32.8	35.1	35.7	37.9
RIPU [40]	19.4	27.3	32.7	34.5	36.2	38.7
S4AL [30]	23.6	30.8	32.4	35.9	38.8	39.7
PBAL [28]	15.9	27.1	31.9	35.8	38.7	39.6
AcZeroTS (ours)	**33.2**	**35.9**	**36.8**	**38.1**	**40.0**	**42.3**

Table 3. Comparison of AcZeroTS with previous methods on Pascal VOC 2012. 'C' represents the active learning cycle.

Method	C1	C2	C3	C4	C5	C6
Random	45.7	49.8	55.8	62.1	74.2	77.5
Entropy [12]	46.2	50.4	57.0	64.1	78.1	80.4
Coreset [31]	47.7	53.0	60.9	71.9	79.1	83.4
DEAL [41]	45.2	50.1	57.6	66.8	77.6	81.6
MADA [27]	46.8	51.3	58.1	68.5	79.0	82.1
RIPU [40]	45.6	50.9	58.0	69.9	79.3	83.2
S4AL [30]	48.1	53.4	57.9	73.0	80.1	84.7
PBAL [28]	45.4	51.2	59.7	70.9	80.0	83.9
AcZeroTS (ours)	**49.1**	**54.3**	**62.6**	**77.1**	**83.5**	**86.9**

Comparison Method. We compare AcZeroTS with following 8 methods: Random, Entropy [12], Coreset [31], DEAL [41], MADA [27], RIPU [40], S4AL [30], and PBAL [28]. More descriptions of these comparing methods are presented in *Supplementary Materials* (Sec. 7.2).

4.3. Comparison with State-of-the-art Methods

We compare AcZeroTS with SoTA active learning methods on the TNBC (Tab. 1) and HPBC (Tab. 2) datasets. To further evaluate the effectiveness of our method, we also conduct experiments on a natural image dataset (Pascal VOC 2012) in Tab. 3. The experimental results show that

Table 4. Comparison of ProZS with previous methods on TNBC and HPBC. w2v: word2vec, ft: fasttext. pAcc: pixel-wise classification accuracy. mIoU(S): seen mIoU. mIoU(U): unseen mIoU. hIoU: harmonic mIoU.

Method	Text Encoder	TNBC												HPBC			
		1 unseen				2 unseen				3 unseen				1 unseen			
		pAcc	mIoU(S)	mIoU(U)	hIoU	pAcc	mIoU(S)	mIoU(U)	hIoU	pAcc	mIoU(S)	mIoU(U)	hIoU	pAcc	mIoU(S)	mIoU(U)	hIoU
SPNet [39]	ft [16], w2v [25]	57.4	35.2	26.9	30.5	58.3	41.5	18.2	25.3	56.5	42.9	16.7	24.0	59.7	53.7	17.9	26.9
ZS3 [5]	w2v	59.9	34.1	30.2	32.0	58.8	40.3	24.3	30.3	56.3	40.3	20.2	26.9	60.8	55.1	20.2	29.6
CaGNet [10]	w2v	63.3	39.2	34.1	36.5	60.6	45.2	24.6	31.9	60.2	46.9	23.4	31.2	69.3	59.1	26.4	36.5
SIGN [7]	w2v	60.2	35.4	34.9	35.1	64.5	46.9	25.2	32.8	60.5	48.5	23.8	31.9	68.5	58.7	20.3	30.2
Joint [3]	w2v	64.7	41.7	39.4	40.5	65.6	48.4	27.4	35.0	62.9	51.9	26.1	34.7	69.6	59.5	22.4	32.5
ZegFormer [8]	PLIP [14]	73.7	49.1	48.5	48.8	72.9	55.3	29.9	38.8	71.9	55.3	27.6	36.8	71.5	61.1	29.8	40.1
ZegCLIP [47]	PLIP	74.6	50.6	48.9	49.7	73.2	55.9	29.5	38.6	72.4	55.6	28.3	37.5	71.4	61.0	31.5	41.5
SaLIP [1]	PLIP	70.0	42.7	41.8	42.2	68.9	51.2	25.9	34.4	69.7	53.5	26.4	35.4	70.2	60.7	27.6	37.9
USE [36]	PLIP	74.5	51.8	49.3	50.5	74.2	57.3	28.1	37.7	72.6	56.0	28.6	37.9	71.9	61.4	31.5	41.6
TagCLIP [20]	PLIP	74.8	53.1	50.4	51.7	73.5	57.4	30.1	39.5	72.8	56.2	28.9	38.2	72.2	61.8	32.1	42.3
ProZS (ours)	PLIP	**76.5**	**58.7**	**52.8**	**55.6**	**74.5**	**58.8**	**33.7**	**42.8**	**73.1**	**56.9**	**30.5**	**39.7**	**74.2**	**62.1**	**35.1**	**44.9**

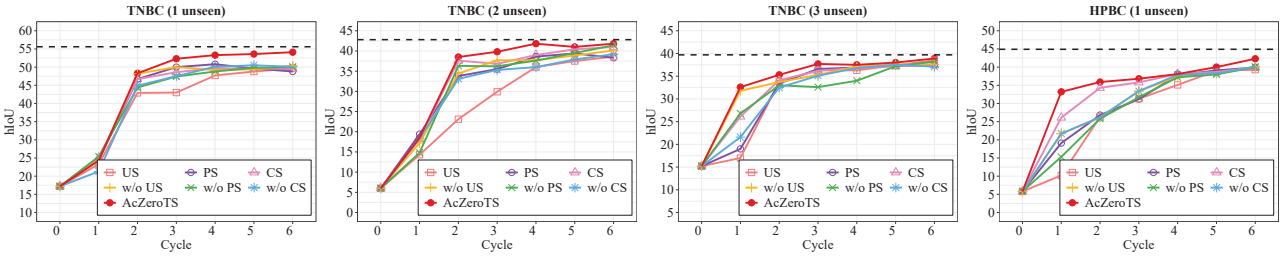


Figure 5. Ablation study for different query strategies. The black dashed line represents the result after labeling all unlabeled images.

AcZeroTS consistently outperforms existing active learning approaches across different datasets while significantly reducing annotation costs. Specifically, at the 3rd cycle of active learning (C3), AcZeroTS achieves 52.3% hIoU on TNBC (1 unseen class), surpassing the second-best method (46.0%, RIPU) by 6.3%. For 2 unseen classes, AcZeroTS reaches 39.8%, outperforming the next-best method (36.8%, MADA) by 3.0%, and for 3 unseen classes, it scores 37.7%, a 4.6% improvement over the second-best result (33.1%, S4AL). A similar trend is observed on the HPBC dataset, where AcZeroTS achieves 36.8% hIoU at C3, exceeding the second-best method (32.8%, MADA) by 4.0%. On the Pascal VOC 2012 dataset, AcZeroTS reaches 77.1% hIoU at C4, outperforming the second-best method (73.0%, S4AL) by 4.1%. These improvements arise from a key limitation in existing active learning methods, which are primarily designed for fully supervised segmentation and overlook the challenge of selecting valuable samples in the presence of unseen classes. In contrast, AcZeroTS integrates uncertainty, diversity, and prototype-based selection criteria, ensuring that sample selection not only improves segmentation performance on seen classes but also enhances generalization to unseen categories.

4.4. Ablation Analysis

We analyze the effects of each query strategy in AcZeroTS on the TNBC and HPBC datasets as shown in Fig. 5, where US, PS, CS refer to using uncertainty score, prototype score, and contrast score individually for sample selection. Additionally, we evaluate w/o US, w/o PS, and w/o CS, which examine the impact of removing uncertainty, prototype, or contrast score from the query process. The results indicate that using all three criteria together consistently achieves the best segmentation performance across all cycles and datasets. Specifically, relying on a single selection strategy (US, PS, or CS) leads to suboptimal results, confirming that a multi-faceted query strategy is essential for effective active learning in ZSS. Furthermore, removing any component causes a noticeable performance drop, particularly in the early active learning cycles, highlighting the role of each strategy. The w/o US variant shows that uncertainty-based selection is crucial for identifying informative samples. The w/o PS variant demonstrates that prototype-based selection enhances generalization to unseen classes. Similarly, w/o CS reveals that contrast selection is essential for ensuring sample diversity and refining feature representations. These findings emphasize that the combination of all three query strategies in AcZeroTS is greater than the sum of its parts, ensuring both improved sample selection and enhanced generalization to unseen classes while reducing annotation costs.

4.5. Visualization

We present sample visualization results on TNBC at the 3rd active learning cycle (C3, 35% annotated) in Fig. 6. Compared to SoTA active learning methods (S4AL [30], RIPU [40], MADA [27]), AcZeroTS can more effectively detect unseen tissue regions as well as achieve better segmentation

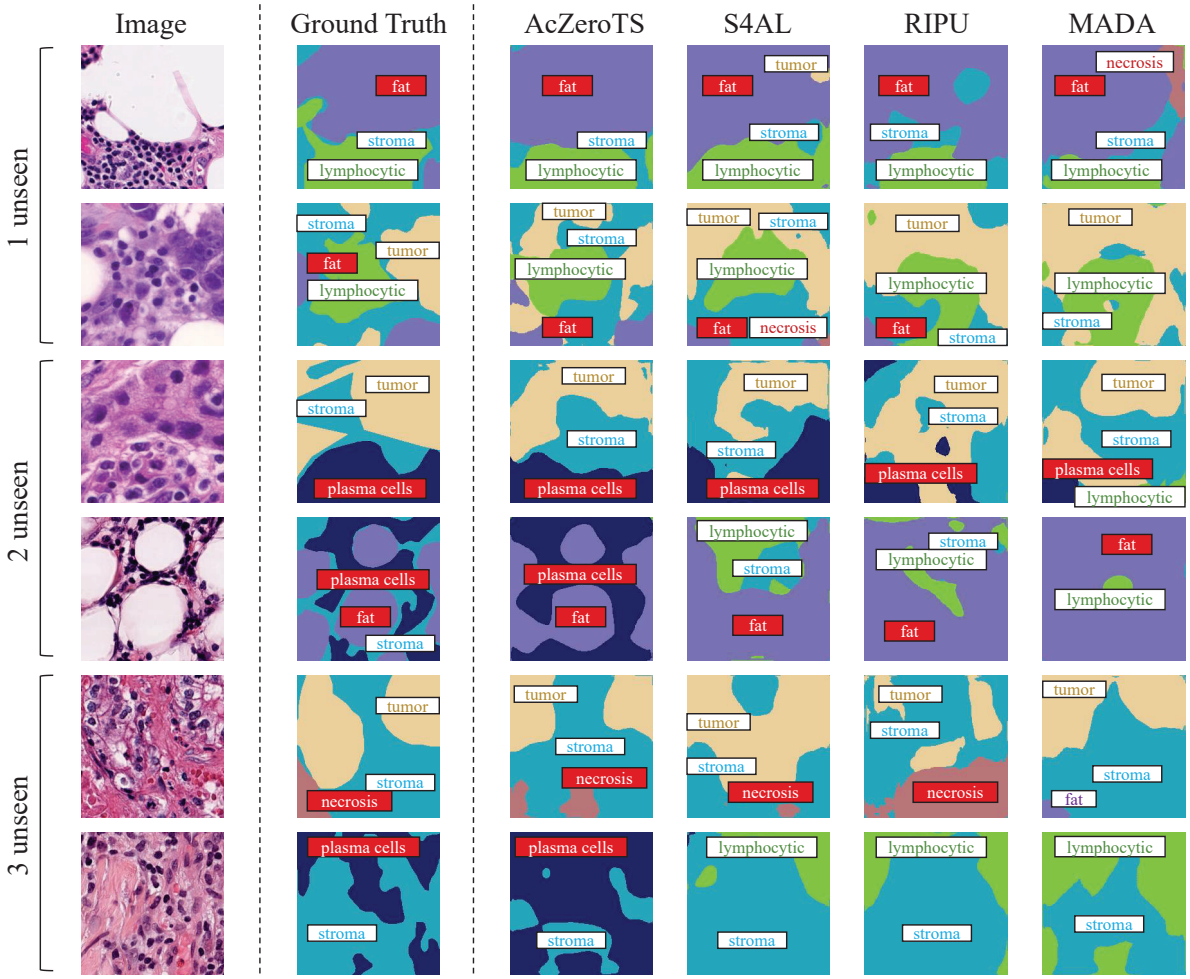


Figure 6. Visualization of segmentation results on TNBC at the 3rd active learning cycle (C3, 35% annotated) under different unseen settings. The white and red tags represent seen and unseen classes separately.

performance on both seen and unseen classes, which further highlights the effectiveness of our framework in reducing annotation efforts while maintaining high segmentation accuracy across both seen and unseen tissue classes.

4.6. Results of ProZS

In this section, we discuss the effectiveness of our ZSS model ProZS. We compare ProZS with SoTA methods (more descriptions are presented in *Supplementary Materials* (Sec. 7.2)) for the ZSS task. From Tab. 4, we can see that ProZS achieves superior performance for tissue segmentation, particularly on the unseen classes. Specifically, for TNBC (2 unseen) and HPBC, the unseen mIoU achieved by the best existing method are 30.1 and 32.1, respectively. ProZS surpasses it by 3.6% on TNBC and 3.0% on HPBC, which clearly showcases the superior generalization ability of ProZS compared to existing approaches.

In *Supplementary Materials*, we further explore the effectiveness of ProZS on the Pascal VOC 2012 dataset (Sec. 8.1). Additionally, we conduct experiments on different tasks (Sec. 8.1), along with parametric analysis (Sec. 8.3), ablation studies on the impact of different text descriptions, different VLMs, and the category-wise pixel-level contrastive loss in ProZS (Sec. 8.2), and an evaluation of model efficiency in terms of parameter count and inference speed (Sec. 8.4).

5. Conclusion

In this paper, we propose AcZeroTS, a novel active learning framework for zero-shot tissue segmentation in pathology images, which is built on a prototype-guided zero-shot segmentation model ProZS. AcZeroTS can predict unseen class tissues with limited annotation budgets, which shows promising potential for detecting diverse tumor microenvironment components in pathology images. The experimental results on two pathology datasets as well as a natural image dataset demonstrate the superior performance of our approach in comparison with the existing studies.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. 62136004, 62272226, 62102188), Key Research and Development Plan of Jiangsu Province, China under Grant BE2022842.

References

[1] Sidra Aleem, Fangyijie Wang, Mayug Maniparambil, Eric Arazo, Julia Dietlmeier, Kathleen Curran, Noel EO' Connor, and Suzanne Little. Test-time adaptation with salip: A cascade of sam and clip for zero-shot medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5184–5193, 2024. 2, 7, 3

[2] Mohamed Amgad, Habiba Elfandy, Hagar Hussein, Lamees A Atteya, Mai AT Elsebaie, Lamia S Abo Elnasr, Rokia A Sakr, Hazem SE Salem, Ahmed F Ismail, Anas M Saad, et al. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics*, 35(18): 3461–3467, 2019. 5

[3] Donghyeon Baek, Youngmin Oh, and Bumsub Ham. Exploiting a joint embedding space for generalized zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9536–9545, 2021. 2, 7, 3

[4] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 6

[5] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 7

[6] Lile Cai, Xun Xu, Jun Hao Liew, and Chuan Sheng Foo. Revisiting superpixels for active learning in semantic segmentation with realistic annotation costs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10988–10997, 2021. 2

[7] Jiaxin Cheng, Soumyaroop Nandi, Prem Natarajan, and Wael Abd-Almageed. Sign: Spatial-information incorporated generative network for generalized zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9556–9566, 2021. 2, 7, 3

[8] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022. 2, 7, 3

[9] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011. 5

[10] Zhangxuan Gu, Siyuan Zhou, Li Niu, Zihan Zhao, and Liqing Zhang. Context-aware feature generation for zero-shot semantic segmentation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1921–1929, 2020. 2, 7

[11] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *International Conference on Learning Representations*, 2020. 4

[12] Alex Holub, Pietro Perona, and Michael C Burl. Entropy-based active learning for object recognition. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2008. 2, 6

[13] Ping Hu, Stan Sclaroff, and Kate Saenko. Uncertainty-aware learning for zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 33:21713–21724, 2020. 2

[14] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023. 3, 7

[15] K Jagadeesh and A Rajendran. Improved model for genetic algorithm-based accurate lung cancer segmentation and classification. *Computer Systems Science & Engineering*, 45(2), 2023. 1

[16] Armand Joulin. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016. 7

[17] Audun Jsang. *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated, 2018. 5

[18] Kwanyoung Kim, Dongwon Park, Kwang In Kim, and Se Young Chun. Task-aware variational adversarial active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8166–8175, 2021. 2

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 3

[20] Jingyao Li, Pengguang Chen, Shengju Qian, Shu Liu, and Jiaya Jia. Tagclip: Improving discrimination ability of zero-shot semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2, 6, 7, 3, 4

[21] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024. 5

[22] Wenjie Luo, Alex Schwing, and Raquel Urtasun. Latent structured active learning. *Advances in neural information processing systems*, 26, 2013. 4

[23] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 3

[24] Christian Mattjie, Luis Vinicius De Moura, Rafaela Ravazio, Lucas Kupssinskü, Otávio Parraga, Marcelo Mussi Delucis, and Rodrigo C Barros. Zero-shot performance of the segment anything model (sam) in 2d medical imaging: A comprehensive evaluation and practical guidelines. In *2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 108–112. IEEE, 2023. 2

[25] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and

phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013. 7

[26] Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In *international conference on machine learning*, pages 7034–7044. PMLR, 2020. 4

[27] Munan Ning, Donghuan Lu, Dong Wei, Cheng Bian, Chenglang Yuan, Shuang Yu, Kai Ma, and Yefeng Zheng. Multi-anchor active domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9112–9122, 2021. 3, 6, 7, 2

[28] Junkun Peng, Mingjie Sun, Eng Gee Lim, Qiufeng Wang, and Jimin Xiao. Prototype guided pseudo labeling and perturbation-based active learning for domain adaptive semantic segmentation. *Pattern Recognition*, 148:110203, 2024. 6, 2

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[30] Aneesh Rangnekar, Christopher Kanan, and Matthew Hoffman. Semantic segmentation with active semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5966–5977, 2023. 6, 7, 2

[31] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *International Conference on Learning Representations (ICLR)*, 2017. 3, 6, 2

[32] Peilun Shi, Jianing Qiu, Sai Mu Dalike Abaxi, Hao Wei, Frank P-W Lo, and Wu Yuan. Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation. *Diagnostics*, 13(11):1947, 2023. 2

[33] Yawar Siddiqui, Julien Valentin, and Matthias Nießner. Viewal: Active learning with viewpoint entropy for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9433–9443, 2020. 2

[34] Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67:101813, 2021. 2

[35] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017. 4

[36] Xiaoqi Wang, Wenbin He, Xiwei Xuan, Clint Sebastian, Jorge Piazentin Ono, Xin Li, Sima Behpour, Thang Doan, Liang Gou, Han-Wei Shen, et al. Use: Universal segment embeddings for open-vocabulary image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4187–4196, 2024. 7, 3, 4

[37] Yawen Wu, Michael Cheng, Shuo Huang, Zongxiang Pei, Yingli Zuo, Jianxin Liu, Kai Yang, Qi Zhu, Jie Zhang, Honghai Hong, et al. Recent advances of deep learning for computational histopathology: principles and applications. *Cancers*, 14(5):1199, 2022. 1

[38] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018. 6

[39] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8256–8265, 2019. 2, 3, 7

[40] Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, and Xinjing Cheng. Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8068–8078, 2022. 2, 6, 7

[41] Shuai Xie, Zunlei Feng, Ying Chen, Songtao Sun, Chao Ma, and Mingli Song. Deal: Difficulty-aware active learning for semantic segmentation. In *Proceedings of the Asian conference on computer vision*, 2020. 2, 6

[42] Fuyong Xing and Lin Yang. Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: a comprehensive review. *IEEE reviews in biomedical engineering*, 9:234–263, 2016. 1

[43] Hongming Xu, Lina Liu, Xiujuan Lei, Mrinal Mandal, and Cheng Lu. An unsupervised method for histological image segmentation based on tissue cluster level graph cut. *Computerized Medical Imaging and Graphics*, 93:101974, 2021. 1

[44] Rui Xu, Zhizhen Wang, Zhenbing Liu, Chu Han, Lixu Yan, Huan Lin, Zeyan Xu, Zhengyun Feng, Changhong Liang, Xin Chen, et al. Histopathological tissue segmentation of lung cancer with bilinear cnn and soft attention. *BioMed Research International*, 2022(1):7966553, 2022. 1

[45] Hui Zhang and Henghui Ding. Prototypical matching and open set rejection for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6974–6983, 2021. 2, 6

[46] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 2, 3

[47] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11175–11185, 2023. 2, 6, 7, 3

[48] Yingli Zuo, Yawen Wu, Zixiao Lu, Qi Zhu, Kun Huang, Daoqiang Zhang, and Wei Shao. Identify consistent imag-

ing genomic biomarkers for characterizing the survival-associated interactions between tumor-infiltrating lymphocytes and tumors. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 222–231. Springer, 2022. 2