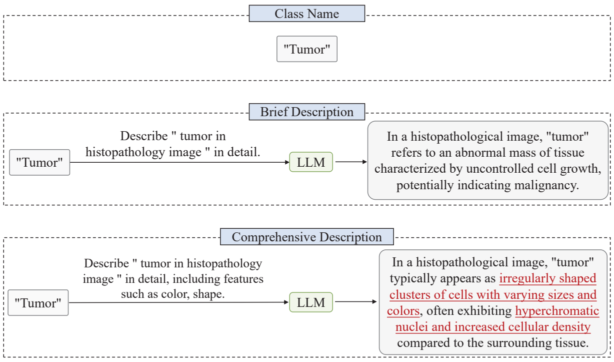
AcZeroTS: Active Learning for Zero-shot Tissue Segmentation in Pathology Images

Supplementary Material



Describe " tumor in histopathology image " in detail.

In a histopathological image, "tumor" refers to an abnormal mass of tissue characterized by uncontrolled cell growth, potentially indicating malignancy.

Describe " tumor in histopathology image " in detail, including features such as color, shape.

In a histopathological image, "tumor" typically appears as irregularly shaped clusters of cells with varying sizes and colors, often exhibiting hyperchromatic nuclei and increased cellular density compared to the surrounding tissue.

Figure 7. Presentation of different description texts.

6. Generated Text Description

In Sec. 3.3, we generate description texts for different tissue categories by considering their appearance attributes using a large language model, rather than relying on generic prompts like “a photo of {}”. Specifically, we use the following prompts: *“Describe {} in the histopathology image in detail, including features such as color, shape.”* / *“Describe {} in the natural image in detail, including features such as color, shape.”* We list the descriptions for each tissue category in TNBC and HPBC datasets as follows:

- “background” refers to the overall appearance of the tissue surrounding specific features of interest. It may appear as a homogeneous field of pale pink or blue.
- “tumor” typically appears as irregularly shaped clusters of cells with varying sizes and colors, often exhibiting hyperchromatic nuclei and increased cellular density compared to the surrounding tissue.
- “stroma” presents as a network of thin, elongated fibers or cells that can be pink or pale in color, providing structural support to the tissue.
- “lymphocytic infiltrate” manifests as small, round, dark-staining lymphocytes scattered throughout the tissue, indicating an immune response with clusters of these cells forming lymphoid aggregates.
- “necrosis or debris” is characterized by areas of amorphous, pink-staining material or cellular remnants, often lacking cellular details.
- “fat” presents as clear, vacuolated spaces with a white or light yellow appearance, indicating the presence of adipose tissue.
- “plasma cells” are typically round cells with eccentric nuclei and a clear perinuclear halo, often indicating an immune response.

Additionally, the detailed descriptions for each category in the natural image dataset (Pascal VOC 2012) are provided below:

- “aeroplane” typically appears as a flying aircraft with wings and a fuselage, often featuring distinct shapes and colors against the sky or clouds.
- “bicycle” is recognizable by its two wheels, frame, handlebars, and pedals, commonly seen on roads or in outdoor settings.
- “bird” is characterized by its feathered body, wings, beak, and often seen perched or in flight.
- “boat” refers to watercraft with a hull for sailing on water, varying in size and design.
- “bottle” is a container typically made of glass or plastic, often used for storing liquids.
- “bus” is a large vehicle designed for transporting passengers, commonly seen on roads in urban areas.
- “car” is a motor vehicle with four wheels used for transportation on roads, varying in make and model.
- “cat” is a domesticated feline animal with whiskers, fur, and often seen lounging or moving about.
- “chair” is a piece of furniture designed for sitting, featuring a seat and back support.
- “cow” is a domesticated bovine animal with horns and distinctive black and white markings.
- “table” is a piece of furniture with a flat top supported by legs, commonly used for various activities.
- “dog” is a domesticated canine animal known for its loyalty and varied breeds.
- “horse” is a large mammal with hooves and a mane, often used for riding or pulling loads.
- “motorbike” is a two-wheeled vehicle powered by a motor, commonly used for transportation.
- “person” refers to a human individual, typically seen engaged in various activities.
- “pottedplant” is a plant cultivated in a pot or container, often found indoors or in gardens.
- “sheep” is a domesticated ruminant animal with woolly fleece.
- “sofa” is a type of seating furniture designed for relaxation, often found in living rooms.
- “train” is a connected series of railway vehicles for transporting passengers or goods.
- “tvmonitor” is a screen for displaying visual content, commonly found in homes or public spaces.

Table 5. Comparison of ProZS with previous methods under the transductive setting on TNBC. ST: self-training, w2v: word2vec, ft: fasttext. pAcc: pixel-wise classification accuracy. mIoU(S): seen mIoU. mIoU(U): unseen mIoU. hIoU: harmonic mIoU.

Method	Text Encoder	1 unseen				2 unseen				3 unseen			
		pAcc	mIoU(S)	mIoU(U)	hIoU	pAcc	mIoU(S)	mIoU(U)	hIoU	pAcc	mIoU(S)	mIoU(U)	hIoU
SPNet+ST [39]	ft, w2v	59.9	34.2	30.9	32.5	60.3	41.4	24.5	30.8	62.1	41.5	25.1	31.3
ZS5 [5]	w2v	60.2	36.4	33.9	35.1	59.3	40.6	29.5	34.2	62.3	40.3	29.2	33.9
CaGNet+ST [10]	w2v	67.8	41.1	39.6	40.3	66.5	47.7	29.8	36.7	67.5	48.5	29.8	36.9
MaskCLIP+ [46]	PLIP	74.7	53.7	49.9	51.7	72.8	55.5	31.4	40.1	73.4	55.8	35.6	43.5
ZegCLIP+ST [47]	PLIP	74.9	51.1	50.3	50.7	74.1	55.1	31.8	40.3	73.5	56.3	35.7	43.7
TagCLIP+ST [20]	PLIP	76.9	53.3	51.1	52.2	74.3	57.8	32.8	41.9	74.1	56.6	35.0	43.3
ProZS+ST (ours)	PLIP	**79.4**	**59.8**	**53.7**	**56.6**	**75.8**	**59.4**	**34.5**	**43.6**	**74.2**	**56.9**	**36.7**	**44.6**

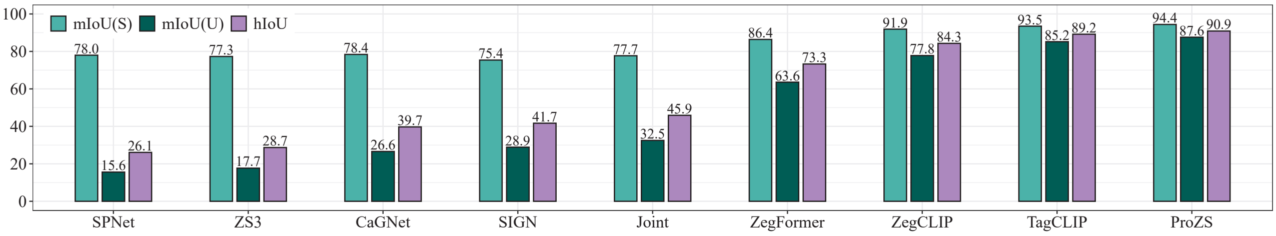


Figure 8. Comparison of ProZS with previous methods on Pascal VOC 2012.

Table 6. Comparison of ProZS with previous methods under the transductive setting on HPBC. ST: self-training, w2v: word2vec, ft: fasttext. pAcc: pixel-wise classification accuracy. mIoU(S): seen mIoU. mIoU(U): unseen mIoU. hIoU: harmonic mIoU.

Method	Text Encoder	pAcc	mIoU(S)	mIoU(U)	hIoU
SPNet+ST [39]	ft, w2v	66.5	58.5	22.1	32.1
ZS5 [5]	w2v	68.6	56.7	24.6	34.3
CaGNet+ST [10]	w2v	70.5	59.0	30.8	40.5
MaskCLIP+ [46]	PLIP	73.4	61.7	33.3	43.3
ZegCLIP+ST [47]	PLIP	72.9	61.8	35.6	45.2
TagCLIP+ST [20]	PLIP	74.8	62.2	36.9	46.3
ProZS+ST (ours)	PLIP	**75.5**	**62.6**	**39.3**	**48.3**

7. Experimental Setting

7.1. Evaluation Metric

In Sec. 4.2, we introduce harmonic mean IoU (hIoU) as the evaluation metric to evaluate the tissue segmentation results among different methods. We list the definition of harmonic mean IoU (hIoU) as follows:

$$\mathrm{hIoU} = \frac{2 \times \mathrm{mIoU(S)} \times \mathrm{mIoU(U)}}{\mathrm{mIoU(S)} + \mathrm{mIoU(U)}}, \tag{12}$$

where $\mathrm{mIoU(S)}$ and $\mathrm{mIoU(U)}$ represent mIoU of seen classes and unseen classes, respectively.

7.2. Comparison Method

In Sec. 4.2 and Sec. 4.6, we compare our AcZeroTS and ProZS with previous methods, the detailed definition are provided below.

7.2.1. Comparison Methods with AcZeroTS

(1) Random: Randomly select samples for annotation.
(2) Entropy [12]: Select samples based on the model entropy.
(3) Coreset [31]: Select samples by the query criteria of representativeness.
(4) DEAL [41]: A difficulty-aware active learning network for semantic segmentation of images.
(5) MADA [27]: Select representative and complimentary samples from the unlabeled pool for annotation.
(6) RIPU [40]: Select samples based on the region impurity for domain adaptive semantic segmentation.
(7) S4AL [30]: Query samples by considering semi-supervision based pseudo labels.
(8) PBAL [28] A prototype-guided pseudo-label generating approach that leverages the relationships between source prototypes and target features for sample querying.

7.2.2. Comparison Methods with ProZS

(1) SPNet [39]: A zero-shot semantic segmentation method using semantic projection network.
(2) ZS3 [5]: An architecture combining rich text and image embeddings for zero-shot segmentation.
(3) CaGNet [10]: A context-aware feature generation method for zero-shot segmentation.

Table 7. Ablation study of different category descriptions. Bri. Des.: Brief Description, Com. Des.: Comprehensive Description.

Description Text	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
Class Label	74.8	57.5	45.5	50.8	73.4	58.4	25.7	35.7	72.4	56.2	27.0	36.5	72.4	62.0	31.7	42.0
Bri. Des.	75.9	57.7	50.2	53.7	74.4	58.6	31.4	40.9	72.8	56.2	29.7	38.9	73.2	61.9	33.9	43.8
Com. Des.	**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**

Table 8. Ablation study of category-wise pixel-level contrastive loss. w/o CPC: without category-wise pixel-level contrastive loss.

Method	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
w/o CPC	76.2	57.6	51.1	54.2	74.5	58.7	32.8	42.1	72.9	56.6	30.0	39.2	74.2	61.9	34.7	44.5
ProZS (ours)	**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**

(4) SIGN [7]: A spatial-information guided generative network for zero-shot semantic segmentation.

(5) Joint [3]:Exploiting the joint embedding of image and text for generalized semantic segmentation.

(6) ZegFormer [8]: A simple but effective zero-shot semantic segmentation model based on the decoupling formulation.

(7) ZegCLIP [47]: A efficient one-stage straightforward zero-shot semantic segmentation method based on the CLIP.

(8) SaLIP [1]: A method integrating SAM and CLIP into a unified framework for medical image segmentation.

(9) USE [36]: An open-vocabulary image segmentation task involves partitioning images into semantically meaningful segments and classifying them with flexible text-defined categories.

(10) TagCLIP [20]: A trusty-aware guided CLIP model for zero-shot segmentation.

(11) MaskCLIP+ [46]:Leverages CLIP to provide pseudo labels for unlabeled pixels, which can be applied to more segmentation-tailored architectures for zero-shot segmentation.

(12) SAM [19]: A vision foundation model for segmenting anything.

(13) MedSAM [23]: A vision foundation model for segmenting medical images.

8. Additional Experiments

8.1. Effectiveness of ProZS

8.1.1. Effectiveness on Natural Dataset

To further validate the effectiveness of our method, we extend ProZS for natural image segmentation on the Pascal VOC 2012, and compare it with SoTA methods by replacing PLIP encoders with CLIP's. We present the mIoU on both seen and unseen classes as well as hIoU in Fig. 8, and the results further demonstrate the superiority of ProZS in tackling ZSS tasks.

Table 9. Comparison of ProZS with previous methods on fully supervised task.

Method	TNBC			HPBC		
	pAcc	mIoU(S)	mIoU(U)	pAcc	mIoU(S)	mIoU(U)
TagCLIP [20]	79.2	54.1	-	80.1	55.9	-
ProZS (ours)	**81.1**	**59.3**	**-**	**80.9**	**57.2**	**-**

Table 10. Cross-dataset segmentation results from HPBC to TNBC.

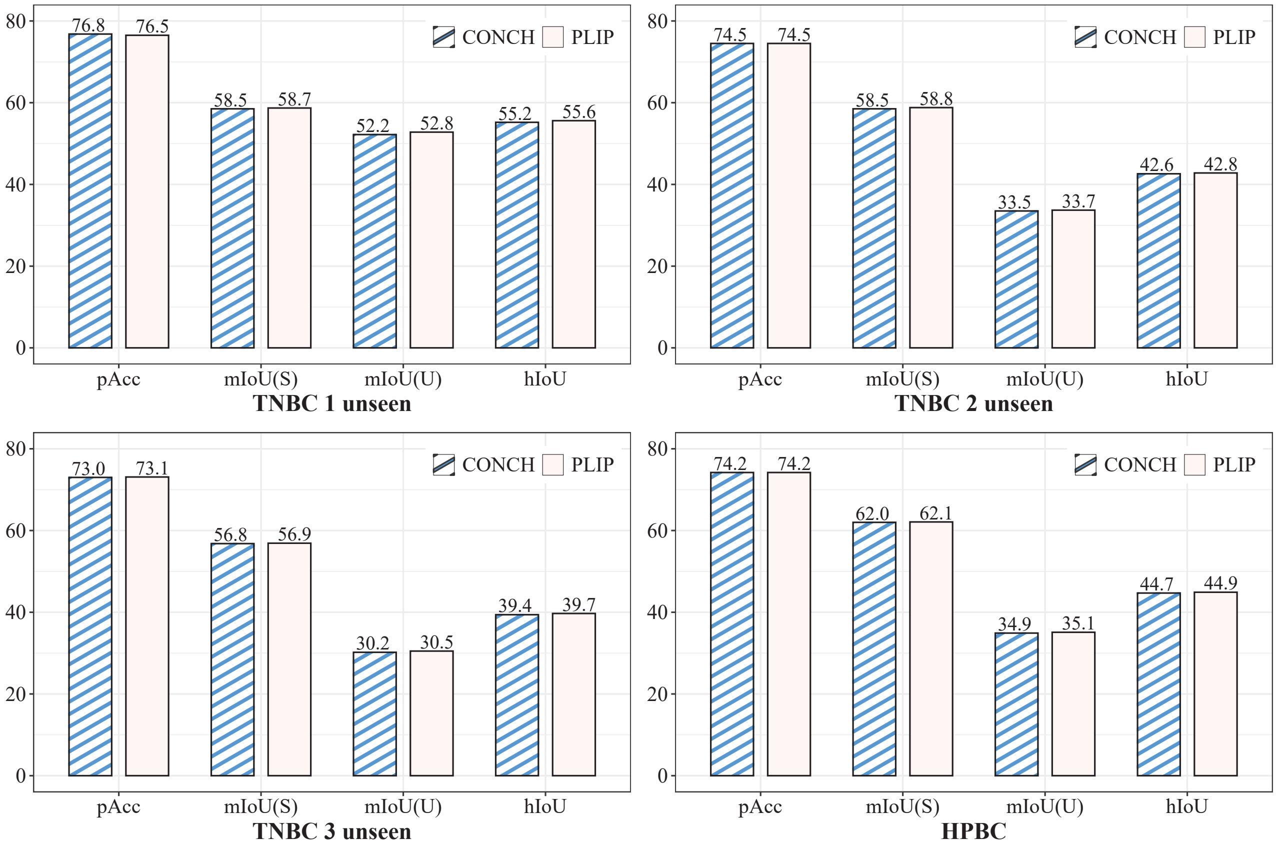
Method	pAcc	mIoU
USE [36]	58.5	29.8
TagCLIP [20]	59.4	30.9
ProZS (ours)	**62.1** (+2.7)	**33.5** (+2.6)

8.1.2. Transductive Task

Generally, GZSS can be divided into two different settings: inductive and transductive. Inductive setting only contains information from seen classes during training. For the transductive setting, partial information (e.g., class name and vision feature) from unseen classes can be accessed in the training process except for the ground truth masks. We evaluate our method under the inductive setting in Sec. 4, and also assess ProZS under the transductive setting on TNBC (Tab. 5) and HPBC (Tab. 6). In the transductive setting, we train ProZS on seen classes for first 50% epochs and then utilize self-training to generate pseudo labels in the remaining training processes. Our results further demonstrate that ProZS can achieve superior segmentation performance for the transductive task. As shown in Tab. 5 and Tab. 6, ProZS not only showcases impressive results on unseen classes but also maintains excellent performance on seen categories for tissue segmentation. More specifically, ProZS outperforms

Table 11. Effects of different regularization parameters λ.

λ	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
0.1	76.2	57.8	52.5	55.0	74.3	58.4	33.0	42.2	73.1	56.1	30.0	39.1	74.2	61.7	35.0	44.7
0.01	**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**
0.001	76.4	57.6	44.8	50.4	74.5	58.0	30.6	40.1	72.9	56.3	29.7	38.9	74.1	61.9	34.4	44.2



TNBC 1 unseen

TNBC 2 unseen

TNBC 3 unseen

HPBC

Figure 9. Ablation study of different vision-language models (VLMs) on TNBC and HPBC.

current SoTA by 2.6% on TNBC and 2.4% on HPBC in mIoU(U).

8.1.3. Fully Supervised Task

We also provide the fully supervised learning results in Tab. 9, which demonstrates that ProZS improves the upper bound of ZSS results (mIoU) by 5.2% on TNBC and 1.3% on HPBC.

8.1.4. Cross-Dataset Task

To further investigate the cross-domain generalization ability of ProZS, we conduct extra experiments that we train the model on the seen classes of HPBC and evaluate its performance on the TNBC. We compare our method with TagCLIP [20] and USE [36] in Tab. 10. The result shows that ProZS outperforms its competitors by 2.7% in pAcc and 2.6% in mIoU, highlighting the satisfied generalization ability of our method across different datasets.

8.2. Ablation of ProZS

8.2.1. Effects of Different Text Descriptions

In ProZS, we design text prompts to describe tissue categories using the LLM. Here, we discuss the effects of applying different prompts and evaluate their performance for tissue segmentation. Specifically, we divide the decription texts using different prompts into three categories (class name (without prompts), brief description, and comprehensive description) shown in Fig. 7. The brief description uses *“Describe* $\{\}$ *in the histopathology image in detail.”* as prompts. For the comprehensive description, we ask LLM to provide more detailed attributes of tissue categories, like *“Describe* $\{\}$ *in the histopathology image in detail, including features such as color, shape.”*. All generated comprehensive descriptions can be found in Sec. 6. From Tab. 7, we observe that the utilization of text descriptions yields significantly better results compared to solely using class labels. Additionally, the comprehensive description provides more detailed and informative cues to the model, allowing it to better understand the characteristics and visual appearance of the target category.

8.2.2. Effects of Category-wise Pixel-level Contrastive Loss

We further discuss the effects of our proposed category-wise pixel-level contrastive (CPC) loss of ProZS in Tab. 8. It can be observed that the CPC loss helps to enhance the segmentation performance on both seen and unseen classes, which demonstrates its advantages in transferring the zero-shot ability of vision-language model from image-level to pixel-level.

8.2.3. Effects of Different Vision-language Models

In ProZS, we utilize the vision-language model (VLM) PLIP to encode both image and text. To further investigate the impact of different VLMs, we conduct experiments using CONCH [21], as shown in Fig. 9. The results indicate that while both PLIP and CONCH exhibit comparable feature extraction and representation capabilities for ZSS task, PLIP demonstrates a slight advantage.

8.3. Parametric Analysis

8.3.1. Discussion on the Regularization Parameter λ

In Eq. (5), we use a regularization parameter $\lambda = 0.01$ to weight the contributions of the segmentation loss $\mathcal{L}_{seg}$ and the category-wise pixel-level contrastive loss $\mathcal{L}_{cpc}$ for the zero-shot segmentation task. Here, we further discuss the effects of λ in the range of $\{0.1, 0.01, 0.001\}$, and the results are shown in Tab. 11. As can be seen from Tab. 11, all indexes fluctuate in a small range on the seen classes. which shows that our method is very robust to the regularization parameter λ for the segmentation tasks on seen tasks. On the other hand, we can observe that the segmentation results on unseen tasks are largely affected with different values of λ, which implies that the selection of λ is very important for segmenting unseen class tissues, and we can derive the best segmentation results if $\lambda = 0.01$.

8.4. Learnable Parameters and Inference Time

We compute the number of learnable parameters and inference time to further evaluate the model’s efficiency. we train ProZS on a single NVIDIA 4090 GPU with 24GB memory, with a total learnable parameter count of 8.17M. The inference time for a single image is approximately 0.009 seconds.