SLPT: Selective Labeling Meets Prompt Tuning on Label-Limited Lesion Segmentation

Fan Bai[1,2,3], Ke Yan[2,3], Xiaoyu Bai[2,3], Xinyu Mao[1], Xiaoli Yin[4], Jingren Zhou[2,3], Yu Shi[4], Le Lu[2], and Max Q.-H. Meng[1,5(✉)]

[1] Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong, China
[2] DAMO Academy, Alibaba Group, Hangzhou, China
[3] Hupan Lab, Hangzhou 310023, China
[4] Department of Radiology, Shengjing Hospital of China Medical University, Shenyang 110004, China
[5] Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen, China
mengqh@sustech.edu.cn

Abstract. Medical image analysis using deep learning is often challenged by limited labeled data and high annotation costs. Fine-tuning the entire network in label-limited scenarios can lead to overfitting and suboptimal performance. Recently, prompt tuning has emerged as a more promising technique that introduces a few additional tunable parameters as prompts to a task-agnostic pre-trained model, and updates only these parameters using supervision from limited labeled data while keeping the pre-trained model unchanged. However, previous work has overlooked the importance of selective labeling in downstream tasks, which aims to select the most valuable downstream samples for annotation to achieve the best performance with minimum annotation cost. To address this, we propose a framework that combines selective labeling with prompt tuning (SLPT) to boost performance in limited labels. Specifically, we introduce a feature-aware prompt updater to guide prompt tuning and a TandEm Selective LAbeling (TESLA) strategy. TESLA includes unsupervised diversity selection and supervised selection using prompt-based uncertainty. In addition, we propose a diversified visual prompt tuning strategy to provide multi-prompt-based discrepant predictions for TESLA. We evaluate our method on liver tumor segmentation and achieve state-of-the-art performance, outperforming traditional fine-tuning with only 6% of tunable parameters, also achieving 94% of full-data performance by labeling only 5% of the data.

Keywords: Active Learning · Prompt Tuning · Segmentation

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43895-0_2.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14221, pp. 14–24, 2023.
https://doi.org/10.1007/978-3-031-43895-0_2

1 Introduction

Deep learning has achieved promising performance in computer-aided diagnosis [1,12,14,24], but it relies on large-scale labeled data to train, which is challenging in medical imaging due to label scarcity and high annotation cost [3,25]. Specifically, expert annotations are required for medical data, which can be costly and time-consuming, especially in tasks such as 3D image segmentation.

Transferring pre-trained models to downstream tasks is an effective solution for addressing the label-limited problem [8], but fine-tuning the full network with small downstream data is prone to overfitting [16]. Recently, prompt tuning [5,18] is emerging from natural language processing (NLP), which introduces additional tunable prompt parameters to the pre-trained model and updates only prompt parameters using supervision signals obtained from a few downstream training samples while keeping the entire pre-trained unchanged. By tuning only a few parameters, prompt tuning makes better use of pre-trained knowledge. It avoids driving the entire model with few downstream data, which enables it to outperform traditional fine-tuning in limited labeled data. Building on the recent success of prompt tuning in NLP [5], instead of designing text prompts and Transformer models, we explore visual prompts on Convolutional Neural Networks (CNNs) and the potential to address data limitations in medical imaging.

However, previous prompt tuning research [18,28], whether on language or visual models, has focused solely on the model-centric approach. For instance, CoOp [29] models a prompt's context using a set of learnable vectors and optimizes it on a few downstream data, without discussing what kind of samples are more suitable for learning prompts. VPT [13] explores prompt tuning with a vision Transformer, and SPM [17] attempts to handle downstream segmentation tasks through prompt tuning on CNNs, which are also model-centric. However, in downstream tasks with limited labeled data, selective labeling as a data-centric method is crucial for determining which samples are valuable for learning, similar to Active Learning (AL) [23]. In AL, given the initial labeled data, the model actively selects a subset of valuable samples for labeling and improves performance with minimum annotation effort. Nevertheless, directly combining prompt tuning with AL presents several problems. First, unlike the task-specific models trained with initial data in AL, the task-agnostic pre-trained model (e.g., trained by related but not identical supervised or self-supervised task) is employed for data selection with prompt tuning. Second, in prompt tuning, the pre-trained model is frozen, which may render some AL methods inapplicable, such as those previously based on backbone gradient [9] and feature [19]. Third, merging prompt tuning with AL takes work. Their interplay must be considered. However, previous AL methods [27] did not consider the existence of prompts or use prompts to estimate sample value.

Therefore, this paper proposes the first framework for selective labeling and prompt tuning (SLPT), combining model-centric and data-centric methods to improve performance in medical label-limited scenarios. We make three main contributions: (1) We design a novel feature-aware prompt updater embedded in the pre-trained model to guide prompt tuning in deep layers. (2) We propose

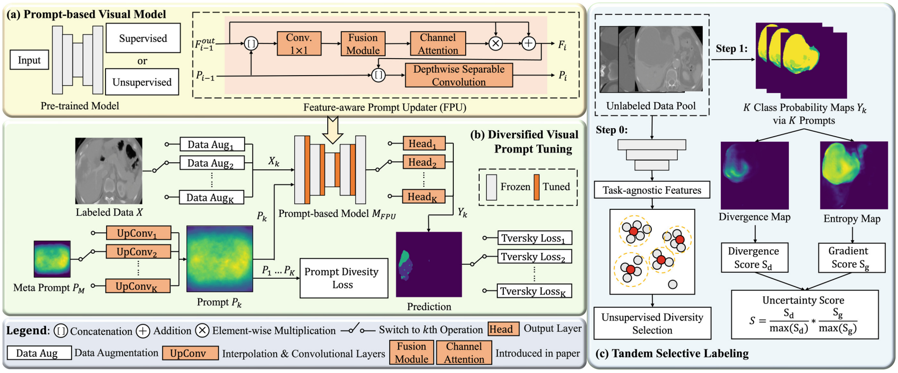
(a) Prompt-based Visual Model
Input
Supervised
or
Unsupervised
Pre-trained Model
F_{i-1}^{out}
P_{i-1}
Conv. 1×1
Fusion Module
Channel Attention
Depthwise Separable Convolution
F_i
P_i
Feature-aware Prompt Updater (FPU)
(b) Diversified Visual Prompt Tuning
Labeled Data X
Data Aug$_1$
Data Aug$_2$
Data Aug$_K$
X_k
Prompt-based Model M_{FPU}
Head$_1$
Head$_2$
Head$_K$
Y_k
Frozen
Tuned
Meta Prompt P_M
UpConv$_1$
UpConv$_2$
UpConv$_K$
Prompt P_k
P_k
$P_1 \ldots P_K$
Prompt Divesity Loss
Prediction
Tversky Loss$_1$
Tversky Loss$_2$
Tversky Loss$_K$
Legend: [] Concatenation ⊕ Addition ⊗ Element-wise Multiplication Switch to kth Operation
Head Output Layer
Data Aug Data Augmentation
UpConv Interpolation & Convolutional Layers
Fusion Module Channel Attention Introduced in paper
Unlabeled Data Pool
Step 1:
K Class Probability Maps Y_k via K Prompts
Step 0:
Task-agnostic Features
Unsupervised Diversity Selection
Divergence Map
Entropy Map
Divergence Score $\mathrm{S_d}$
Gradient Score $\mathrm{S_g}$
Uncertainty Score
$S = \frac{\mathrm{S_d}}{\max(\mathrm{S_d})} * \frac{\mathrm{S_g}}{\max(\mathrm{S_g})}$
(c) Tandem Selective Labeling

Fig. 1. Workflow of SLPT: **(1)** Create an initial label set via the pre-trained model for unsupervised diversity selection (subplot c step 0). **(2)** Insert a feature-aware prompt updater (subplot a) into the pre-trained model for prompt tuning with initial labels. **(3)** Use diversified visual prompt tuning (subplot b) to obtain prompt-based discrepant predictions. **(4)** Select valuable data by prompt-based uncertainty (subplot c step 1) and update the prompt-based model accordingly. Note: The orange modules are tunable for prompt tuning, while the gray ones are frozen. Please zoom in for details.

a diversified visual prompt tuning mechanism that provides multi-prompt-based discrepant predictions for selective labeling. (3) We introduce the TESLA strategy which includes both unsupervised diversity selection via task-agnostic features and supervised selection considering prompt-based uncertainty. The results show that SLPT outperforms fine-tuning with just 6% of tunable parameters and achieves 94% of full-data performance by selecting only 5% of labeled data.

2 Methodology

Given a task-agnostic pre-trained model and unlabeled data for an initial medical task, we propose SLPT to improve model performance. SLPT consists of three components, as illustrated in Fig. 1: (a) a prompt-based visual model, (b) diversified visual prompt tuning, and (c) tandem selective labeling. Specifically, with SLPT, we can select valuable data to label and tune the model via prompts, which helps the model overcome label-limited medical scenarios.

2.1 Prompt-Based Visual Model

The pre-trained model, learned by supervised or unsupervised training, is a powerful tool for improving performance on label-limited downstream tasks. Fine-tuning a large pre-trained model with limited data may be suboptimal and prone to overfitting [16]. To overcome this issue, we draw inspiration from NLP [18] and explore prompt tuning on visual models. In order to facilitate prompt tuning on the model's deep layers, we introduce the Feature-aware Prompt Updater (FPU). FPUs are inserted into the network to update deep prompts and features. In Fig. 1(a), an FPU receives two inputs, feature map F_{i-1}^{out} and prompt P_{i-1}, of the same shape, and updates to F_i and P_i through two parallel branches. In the feature branch, F_{i-1}^{out} and P_{i-1} are concatenated and fed into a 1x1 convolution and fusion module. The fusion module utilizes ASPP [7] to extract multi-scale contexts. Then a SE [11] module for channel attention enhances context by channel. Finally, the attention output and F_{i-1}^{out} are element-wise multiplied and added to obtain the updated feature F_i. In the prompt branch, the updated feature F_i is concatenated with the previous prompt P_{i-1}, and a parameter-efficient depth-separable convolution is employed to generate the updated prompt P_i.

To incorporate FPU into a pre-trained model, we consider the model comprising N modular M_i $(i = 1, ..., N)$ and a head output layer. After each M_i, we insert an FPU_i. Given the input F_{i-1}^{in} and prompt P_{i-1}, we have the output feature F_i, updated prompt P_i and prediction Y as follows:

$$F_{i-1}^{out} = M_i(F_{i-1}^{in}), \quad F_i, P_i = \mathrm{FPU}_i(F_{i-1}^{out}, P_{i-1}), \quad Y = \mathrm{Head}(F_N) \tag{1}$$

where input $X = F_0$, FPU and Head are tuned while M_i is not tunable.

2.2 Diversified Visual Prompt Tuning

Inspired by multi-prompt learning [18] in NLP, we investigate using multiple visual prompts to evaluate prompt-based uncertainty. However, initializing and optimizing K prompts directly can significantly increase parameters and may not ensure prompt diversity. To address these challenges, we propose a diversified visual prompt tuning approach. As shown in Fig. 1(b), our method generates K prompts $P_k \in \mathbb{R}^{1\times D\times H\times W}$ from a meta prompt $P_M \in \mathbb{R}^{1\times\frac{D}{2}\times\frac{H}{2}\times\frac{W}{2}}$ through K different upsampling and convolution operations $UpConv_k$. P_M is initialized from the statistical probability map of the foreground category, similar to [17]. Specifically, we set the foreground to 1 and the background to 0 in the ground-truth mask, and then average all masks and downsample to $1 \times \frac{D}{2} \times \frac{H}{2} \times \frac{W}{2}$. To enhance prompt diversity, we introduce a prompt diversity loss L_{div} that regularizes the cosine similarity between the generated prompts and maximizes their diversity. This loss is formulated as follows:

$$L_{div} = \sum_{k_1=1}^{K-1} \sum_{k_2=k_1+1}^{K} \frac{P_{k_1} \cdot P_{k_2}}{||P_{k_1}||_2 \cdot ||Pk_2||_2} \tag{2}$$

where P_{k_1} and P_{k_2} represent the k_1-th and k_2-th generated prompts, respectively, and $||\cdot||_2$ denotes the L2 norm. By incorporating the prompt diversity loss, we aim to generate a set of diverse prompts for our visual model.

In NLP, using multiple prompts can produce discrepant predictions [2] that help estimate prompt-based uncertainty. Drawing inspiration, we propose a visual prompt tuning approach that associates diverse prompts with discrepant predictions. To achieve this, we design K different data augmentation, heads, and losses based on corresponding K prompts. By varying hyperparameters, we can achieve different data augmentation strengths, increasing the model's diversity and generalization. Different predictions Y_k are generated by K heads, each supervised with a Tversky loss [21] $TL_k = \frac{TP}{TP+\alpha_k FP+\beta_k FN}$, where TP, FP, and FN represent true positive, false positive, and false negative, respectively. To obtain diverse predictions with false positives and negatives, we use different α_k and β_k values in TL_k. The process is formulated as follows:

$$P_k = UpConv_k(P_M), \quad X_k = DA_k(X), \quad Y_k = Head_k(M_{FPU}(X_k, P_k)) \tag{3}$$

$$L = \sum_{k=1}^{K}(\lambda_1 * TL_k(Y_k,^{Y}) + \lambda_2 * CE(Y_k,^{Y})) + \lambda_3 * L_{div} \tag{4}$$

where $k = 1, ..., K$, M_{FPU} is the pre-trained model with FPU, CE is the cross-entropy loss, and $\lambda_1 = \lambda_2 = \lambda_3 = 1$ weight each loss component. ^{Y} represents the ground truth and L is the total loss.

2.3 Tandem Selective Labeling

Previous studies overlook the critical issue of data selection for downstream tasks, especially when available labels are limited. To address this challenge, we propose a novel strategy called TESLA. TESLA consists of two tandem steps: unsupervised diversity selection and supervised uncertainty selection. The first step aims to maximize the diversity of the selected data, while the second step aims to select the most uncertain samples based on diverse prompts.

Step 0: Unsupervised Diversity Selection. Since we do not have any labels in the initial and our pre-trained model is task-agnostic, we select diverse samples to cover the entire dataset. To achieve this, we leverage the pre-trained model to obtain feature representations for all unlabeled data. Although these features are task-independent, they capture the underlying relationships, with similar samples having closer feature distances. We apply the k-center method from Coreset [22], which identifies the B samples that best represent the diversity of the data based on these features. These selected samples are then annotated and serve as the initial dataset for downstream tasks.

Step 1: Supervised Uncertainty Selection. After prompt tuning with the initial dataset, we obtain a task-specific model that can be used to evaluate data value under supervised training. Since only prompt-related parameters can be tuned while others are frozen, we assess prompt-based uncertainty via diverse prompts, considering inter-prompts uncertainty and intra-prompts uncertainty. In the former, we compute the multi-prompt-based divergence map D, given K probability predictions Y_k through K diverse prompts P_k, as follows:

$$D = \sum_{k=1}^{K} \mathrm{KL}(Y_k || Y_{\mathrm{mean}}), \quad Y_{\mathrm{mean}} = \frac{1}{K}\sum_{k=1}^{K} Y_k \tag{5}$$

where KL refers to the KL divergence [15]. Then, we have the divergence score $S_d = \mathrm{Mean}(D)$, which reflects inter-prompts uncertainty.

In the latter, we evaluate intra-prompts uncertainty by computing the mean prediction of the prompts and propose to estimate prompt-based gradients as the model's performance depends on the update of prompt parameters θ_p. However, for these unlabeled samples, computing their supervised loss and gradient directly is not feasible. Therefore, we use the entropy of the model's predictions as a proxy for loss. Specifically, we calculate the entropy-based prompt gradient score S_g for each unlabeled sample as follows:

$$S_g = \sum_{\theta_p} ||\nabla_{\theta_p}(-\sum Y_{mean} * \log Y_{mean})||_2 \tag{6}$$

To avoid manual weight adjustment, we employ multiplication instead of addition. We calculate our uncertainty score S as follows:

$$S = \frac{S_d}{\max(S_d)} \times \frac{S_g}{\max(S_g)} \tag{7}$$

where $\max(\cdot)$ finds the maximum value. We sort the unlabeled data by their corresponding S values in ascending order and select the top B data to annotate.

3 Experiments and Results

3.1 Experimental Settings

Datasets and Pre-trained Model. We conducted experiments on automating liver tumor segmentation in contrast-enhanced CT scans, a crucial task in liver cancer diagnosis and surgical planning [1]. Although there are publicly available liver tumor datasets [1,24], they only contain major tumor types and differ in image characteristics and label distribution from our hospital's data. Deploying a model trained from public data to our hospital directly will be problematic. Collecting large-scale data from our hospital and training a new model will be expensive. Therefore, we can use the model trained from them as a starting point and use SLPT to adapt it to our hospital with minimum cost. We collected a dataset from our in-house hospital comprising 941 CT scans with eight categories: hepatocellular carcinoma, cholangioma, metastasis, hepatoblastoma, hemangioma, focal nodular hyperplasia, cyst, and others. It covers both major and rare tumor types. Our objective is to segment all types of lesions accurately. We utilized a pre-trained model for liver segmentation using supervised learning on two public datasets [24] with no data overlap with our downstream task. The nnUNet [12] was used to preprocess and sample the data into $24 \times 256 \times 256$ patches for training. To evaluate the performance, we employed a 5-fold cross-validation (752 for selection, 189 for test).

Metrics. We evaluated lesion segmentation performance using pixel-wise and lesion-wise metrics. For pixel-wise evaluation, we used the Dice per case, a commonly used metric [1]. For lesion-wise evaluation, we first do connected component analysis to predicted and ground truth masks to extract lesion instances, and then compute precision and recall per case [20]. A predicted lesion is regarded as a TP if its overlap with ground truth is higher than 0.2 in Dice.

Table 1. Evaluation of different tunings on the lesion segmentation with limited data (40 class-balanced patients). Prec. and Rec. denote precision and recall.

Method	Tuning Type	Trainable Parameters	Pixel-wise			Lesion-wise		Mean
			Dice	Prec	Rec	Prec	Rec	
Fine-tuning	All	44.81M	64.43	**87.69**	59.86	50.84	54.14	63.39
Learn-from-Scratch		44.81M	54.15	73.33	50.25	45.84	45.78	53.87
Encoder-tuning	Part	19.48M	65.61	82.00	61.96	29.36	41.10	56.00
Decoder-tuning		23.64M	67.87	77.96	**70.56**	30.82	35.92	56.63
Head-tuning		0.10M	56.73	74.45	55.57	23.29	29.74	47.96
SPM [17]	Prompt	3.15M	68.60	83.07	69.02	62.15	55.19	67.61
Ours		**2.71M**	**68.76**	79.63	69.76	**64.63**	**61.18**	**68.79**

Competing Approaches. In the prompt tuning experiment, we compared our method with three types of tuning: full parameter update (Fine-tuning, Learn-from-Scratch), partial parameter update (Head-tuning, Encoder-tuning, Decoder-tuning), and prompt update (SPM [17]). In the unsupervised diversity selection experiment, we compared our method with random sampling. In the supervised uncertainty selection experiment, we compared our method with random sampling, diversity sampling (Coreset [22], CoreCGN [6]), and uncertainty sampling (Entropy, MC Dropout [10], Ensemble [4], UncertainGCN [6], Ent-gn [26]). Unlike Ensemble, our method was on multi-prompt-based heads. Furthermore, unlike Ent-gn, which computed the entropy-based gradient from a single prediction, we calculated a stable entropy from the muti-prompt-based mean predictions and solely considered the prompt gradient.

Training Setup. We conducted the experiments using the Pytorch framework on a single NVIDIA Tesla V100 GPU. The nnUNet [12] framework was used for 3D lesion segmentation with training 500 epochs at an initial learning rate of 0.01. We integrated 13 FPUs behind each upsampling or downsampling of nnUNet, adding only 2.7M parameters. During training, we set k = 3 and employed diverse data augmentation techniques such as scale, elastic, rotation, and mirror. Three sets of TL parameters is ($\alpha_{1,2,3} = 0.5, 0.7, 0.3$, $\beta_{1,2,3} = 0.5, 0.3, 0.7$). To ensure fairness and eliminate model ensemble effects, we only used the model's prediction with $k = 1$ during testing. We used fixed random seeds and 5-fold cross-validation for all segmentation experiments.

3.2 Results

Evaluation of Prompt Tuning. Since we aim to evaluate the efficacy of prompt tuning on limited labeled data in Table 1, we create a sub-dataset of approximately 5% (40/752) from the original dataset. Specifically, we calculate the class probability distribution vector for each sample based on the pixel class in the mask and use CoreSet with these vectors to select 40 class-balanced samples. Using this sub-dataset, we evaluated various tuning methods for limited

Table 2. Comparison of data selection methods for label-limited lesion segmentation. Step 0: unsupervised diversity selection. Step 1: supervised uncertainty selection. The labeling budget for each step is 20 patients. Step $+\infty$ refers to fully labeled 752 data.

Step	Method	Pixel-wise			Lesion-wise		Mean
		Dice	Prec	Rec	Prec	Rec	
0	Random	65.58	80.00	65.21	23.46	39.94	54.84
	Ours	68.20	78.97	69.15	32.51	34.67	56.70
1	Random	66.67	79.95	70.67	41.45	39.45	59.64
	Entropy	66.39	80.85	66.96	37.40	39.47	58.21
	MC Dropout [10]	69.23	79.61	69.48	30.43	36.29	57.01
	Ensemble [4]	69.79	80.25	69.54	**64.38**	58.34	68.46
	CoreSet [22]	70.72	79.34	72.03	46.03	51.24	63.87
	CoreGCN [6]	70.91	77.56	72.37	51.73	49.88	64.49
	UncertainGCN [6]	71.44	75.07	**75.62**	72.83	44.99	67.99
	Ent-gn [26]	70.54	79.91	71.42	61.12	56.37	67.87
	Ours (w/o S_d)	69.54	81.97	68.59	60.47	59.82	68.08
	Ours (w/o S_g)	71.01	80.68	69.83	59.42	58.78	67.94
	Ours	**72.07**	**82.07**	72.37	61.21	**61.90**	**69.92**
$+\infty$	Fine-tuning with Full Labeled Data	77.44	85.44	77.15	62.78	68.56	74.27

medical lesion diagnosis data. The results are summarized in Table 1. Fine-tuning all parameters served as the strongest baseline, but our method, which utilizes only 6% tunable parameters, outperformed it by 5.4%. Although SPM also outperforms fine-tuning, our methods outperform SPM by 1.18% and save 0.44M tunable parameters with more efficient FPU. In cases of limited data, fine-tuning tends to overfit on a larger number of parameters, while prompt tuning does not. The pre-trained model is crucial for downstream tasks with limited data, as it improves performance by 9.52% compared to Learn-from-Scratch. Among the three partial tuning methods, the number of tuning parameters positively correlates with the model's performance, but they are challenging to surpass fine-tuning.

Evaluation of Selective Labeling. We conducted steps 0 (unsupervised selection) and 1 (supervised selection) from the unlabeled 752 data and compared our approach with other competing methods, as shown in Table 2. In step 0, without any labeled data, our diversity selection outperformed the random baseline by 1.86%. Building upon the 20 data points selected by our method in step 0, we proceeded to step 1, where we compared our method with eight other data selection strategies in supervised mode. As a result, our approach outperformed other methods because of prompt-based uncertainty, such as Ent-gn and Ensemble, by 2.05% and 1.46%, respectively. Our approach outperformed Coreset by 6.05% and CoreGCN by 5.43%. We also outperformed UncertainGCN by 1.93%. MC Dropout and Entropy underperformed in our prompt tuning, likely due to the difficulty of learning such uncertain data with only a few prompt parameters. Notably, our method outperformed random sampling by 10.28%. These results demonstrate the effectiveness of our data selection approach in practical tasks.

Ablation Studies. We conducted ablation studies on S_d and S_g in TESLA. As shown in Table 2, the complete TESLA achieved the best performance, outperforming the version without S_d by 1.84% and the version without S_g by 1.98%. It shows that each component plays a critical role in improving performance.

4 Conclusions

We proposed a pipeline called SLPT that enhances model performance in label-limited scenarios. With only 6% of tunable prompt parameters, SLPT outperforms fine-tuning due to the feature-aware prompt updater. Moreover, we presented a diversified visual prompt tuning and a TESLA strategy that combines unsupervised and supervised selection to build annotated datasets for downstream tasks. SLPT pipeline is a promising solution for practical medical tasks with limited data, providing good performance, few tunable parameters, and low labeling costs. Future work can explore the potential of SLPT in other domains.

Acknowledgements. The work was supported by Alibaba Research Intern Program. Fan Bai and Max Q.-H. Meng were supported by National Key R&D program of China with Grant No. 2019YFB1312400, Hong Kong RGC CRF grant C4063-18G, and Hong Kong Health and Medical Research Fund (HMRF) under Grant 06171066. Xiaoli Yin and Yu Shi were supported by National Natural Science Foundation of China (82071885).

References

1. Bilic, P., et al.: The liver tumor segmentation benchmark (LiTS). Med. Image Anal. **84**, 102680 (2023)
2. Allingham, J.U., et al.: A simple zero-shot prompt weighting technique to improve prompt ensembling in text-image models. arXiv preprint arXiv:2302.06235 (2023)
3. Bai, F., Xing, X., Shen, Y., Ma, H., Meng, M.Q.H.: Discrepancy-based active learning for weakly supervised bleeding segmentation in wireless capsule endoscopy images. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13438, pp. 24–34. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16452-1_3
4. Beluch, W.H., Genewein, T., Nürnberger, A., Köhler, J.M.: The power of ensembles for active learning in image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9368–9377 (2018)

5. Brown, T.B., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems, pp. 1876–1901 (2020)
6. Caramalau, R., Bhattarai, B., Kim, T.K.: Sequential graph convolutional network for active learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9583–9592 (2021)
7. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
8. Cheplygina, V., de Bruijne, M., Pluim, J.P.: Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med. Image Anal. **54**, 280–296 (2019)
9. Dai, C., et al.: Suggestive annotation of brain tumour images with gradient-guided sampling. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12264, pp. 156–165. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59719-1_16
10. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: International Conference on Machine Learning, pp. 1050–1059. PMLR (2016)
11. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
12. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
13. Jia, M., et al.: Visual prompt tuning. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13693, pp. 709–727. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19827-4_41
14. Kim, M., et al.: Deep learning in medical imaging. Neurospine **16**(4), 657 (2019)
15. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**(1), 79–86 (1951)
16. Kumar, A., Raghunathan, A., Jones, R., Ma, T., Liang, P.: Fine-tuning can distort pretrained features and underperform out-of-distribution. arXiv preprint arXiv:2202.10054 (2022)
17. Liu, L., Yu, B.X., Chang, J., Tian, Q., Chen, C.W.: Prompt-matched semantic segmentation. arXiv preprint arXiv:2208.10159 (2022)
18. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput. Surv. **55**(9), 1–35 (2023)
19. Parvaneh, A., Abbasnejad, E., Teney, D., Haffari, G.R., Van Den Hengel, A., Shi, J.Q.: Active learning by feature mixing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12237–12246 (2022)
20. Powers, D.M.: Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061 (2020)
21. Salehi, S.S.M., Erdogmus, D., Gholipour, A.: Tversky loss function for image segmentation using 3D fully convolutional deep networks. In: Wang, Q., Shi, Y., Suk, H.-I., Suzuki, K. (eds.) MLMI 2017. LNCS, vol. 10541, pp. 379–387. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67389-9_44
22. Sener, O., Savarese, S.: Active learning for convolutional neural networks: a core-set approach. arXiv preprint arXiv:1708.00489 (2017)
23. Settles, B.: Active learning literature survey (2009)
24. Simpson, A.L., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019)

25. Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J.N., Wu, Z., Ding, X.: Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation. Med. Image Anal. **63**, 101693 (2020)
26. Wang, T., et al.: Boosting active learning via improving test performance. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 8566–8574 (2022)
27. Zhan, X., Wang, Q., Huang, K.H., Xiong, H., Dou, D., Chan, A.B.: A comparative survey of deep active learning. arXiv preprint arXiv:2203.13450 (2022)
28. Zhao, T., et al.: Prompt design for text classification with transformer-based models. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 2709–2722 (2021)
29. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. Int. J. Comput. Vision **130**(9), 2337–2348 (2022)