MedCLIP-SAM: Bridging Text and Image Towards Universal Medical Image Segmentation

Taha Koleilat[1(✉)], Hojat Asgariandehkordi[1], Hassan Rivaz[1], and Yiming Xiao[2]

[1] Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada
{taha.koleilat,hojat.asgariandehkordi,hassan.rivaz}@concordia.ca
[2] Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada
yiming.xiao@concordia.ca

Abstract. Medical image segmentation of anatomical structures and pathology is crucial in modern clinical diagnosis, disease study, and treatment planning. To date, great progress has been made in deep learning-based segmentation techniques, but most methods still lack data efficiency, generalizability, and interactability. Consequently, the development of new, precise segmentation methods that demand fewer labeled datasets is of utmost importance in medical image analysis. Recently, the emergence of foundation models, such as CLIP and Segment-Anything-Model (SAM), with comprehensive cross-domain representation opened the door for interactive and universal image segmentation. However, exploration of these models for data-efficient medical image segmentation is still limited but is highly necessary. In this paper, we propose a novel framework, called MedCLIP-SAM that combines CLIP and SAM models to generate segmentation of clinical scans using text prompts in both zero-shot and weakly supervised settings. To achieve this, we employed a new Decoupled Hard Negative Noise Contrastive Estimation (DHN-NCE) loss to fine-tune the BiomedCLIP model and the recent gScoreCAM to generate prompts to obtain segmentation masks from SAM in a zero-shot setting. Additionally, we explored the use of zero-shot segmentation labels in a weakly supervised paradigm to improve the segmentation quality further. By extensively testing three diverse segmentation tasks and medical image modalities (breast tumor ultrasound, brain tumor MRI, and lung X-ray), our proposed framework has demonstrated excellent accuracy. Code is available at https://github.com/HealthX-Lab/MedCLIP-SAM.

Keywords: Image segmentation · Foundation models · Zero-shot learning · Weakly Supervised Semantic Segmentation

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
M. G. Linguraru et al. (Eds.): MICCAI 2024, LNCS 15012, pp. 643–653, 2024.
https://doi.org/10.1007/978-3-031-72390-2_60

1 Introduction

With the increasing availability of radiological technologies, there is a pressing need for accurate and efficient medical image segmentation to aid the study, diagnosis, and treatment of various medical conditions [29]. Deep learning (DL) techniques have been established as state-of-the-art in the domain, but current methods often face three major limitations, hindering their widespread clinical adoption. First, the lack of large well-annotated data sets is a major bottleneck for DL model development. Second, the lack of interactivity and interpretability limits the credence of the methods. Lastly, most trained models are task- and contrast/modality-specific with low flexibility. While many self- and weakly supervised methods [4,8,30] have been proposed to tackle training data efficiency and explainable AI (XAI) methods (e.g., uncertainty estimation [19,21] and saliency map [2,3]) are being actively investigated, cross-domain generalization has been a challenge. Recently, the introduction of foundation models, such as the CLIP (Contrastive Language-Image Pre-Training) [26] and SAM (Segment Anything Model) [14] opened the door for interactive and universal medical image segmentation. To date, several groups have endeavored to adapt CLIP and SAM for radiological tasks from natural images, notably the development of BiomedCLIP [33] and MedSAM [22], which were pre-trained on millions of biomedical data. However, more efficient parameter fine-tuning methods can be beneficial to further boost the performance of these foundation models in radiological applications. On the other hand, with a strong interest in SAM, which requires interactive prompts to guide segmentation, a few techniques were proposed to fine-tune SAM without prompts [9,12], generate prompts through Class Activation Map (CAM) from classification tasks [16,17,20], and to refine its output based on weak supervision [7,13,31]. Still at the nascent phase, using foundation models for interactive and universal medical image segmentation necessitates additional investigation and is of significant interest.

To address the aforementioned needs, we present MedCLIP-SAM, a novel framework that leverages BiomedCLIP [33] and SAM [14] for text-prompt-based interactive and universal medical image segmentation in both zero-shot and weakly supervision settings. The contributions of this work are threefold: **First**, we proposed a novel CLIP training/fine-tuning method, called the Decoupled Hard Negative Noise Contrastive Estimation (DHN-NCE). **Second**, we proposed a zero-shot medical segmentation method by combining CLIP and SAM in radiological tasks for the first time. **Lastly**, a weakly-supervised strategy was explored with the attempt to further refine zero-shot segmentation results, and the full proposed technique was extensively validated on three different segmentation tasks and modalities (breast tumor segmentation in ultrasound, brain tumor segmentation in MRI, and lung segmentation in chest X-ray).

2 Methods and Materials

An overview of the proposed MedCLIP-SAM framework is presented in Fig. 1, organized into three distinct stages: BiomedCLIP fine-tuning employing our new

DHN-NCE loss, zero-shot segmentation guided by text-prompts, and weakly supervised segmentation for potential label refinement.

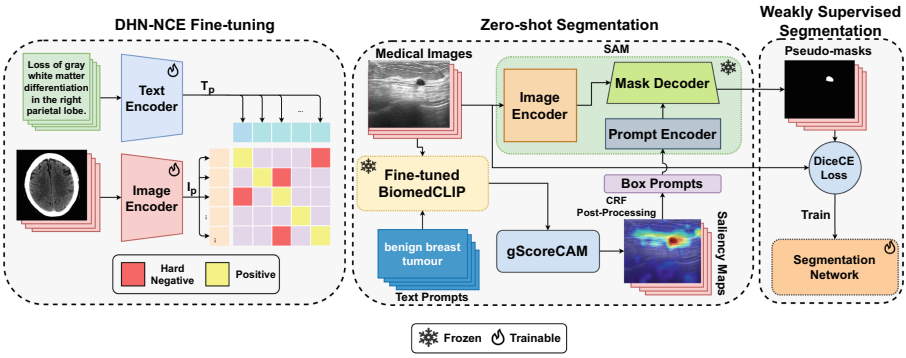


Fig. 1. An overview of the proposed MedCLIP-SAM framework.

2.1 Efficient BiomedCLIP Fine-Tuning with the DHN-NCE Loss

Decoupled Hard Negative Noise Contrastive Estimation Loss. A CLIP model is trained in large datasets of images and the corresponding texts. Specifically, an image encoder and a text encoder are used to extract features of images and texts and project them into vectors of the same dimension, $\mathbf{I}_{p,i}$ and $\mathbf{T}_{p,i}$, respectively. Then, through contrastive learning, an embedding space shared by the image and text vectors is learned so that similar pairs (an image and its description) are closer together and dissimilar ones are farther apart. While BiomedCLIP [33] was trained on medical charts/images and clinical texts, further fine-tuning can effectively benefit medical image-specific tasks. In CLIP training with the conventional InfoNCE loss [23], the *negative-positive-coupling (NPC)* effect [32] can lead to sub-optimal learning efficiency, particularly in small batch sizes while for medical images, more nuanced discrimination between cases within the same imaging categories can be difficult. To solve these, we propose the Decoupled Hard Negative Noise Contrastive Estimation (DHN-NCE) loss, which 1) combines the InfoNCE loss [23] with hard negative sampling [28] to focus on "close samples" and 2) adds decoupling contrastive learning [32] by removing the positive term in the denominator to allow smaller batch sizes. Specifically, the loss function $\mathcal{L}_{DHN-NCE}$ uses weighting functions $(\mathcal{W}^{v\to t}_{\mathbf{I}_{p,i}\mathbf{T}_{p,j}}, \mathcal{W}^{t\to v}_{\mathbf{T}_{p,i}\mathbf{I}_{p,j}})$ to increase the penalty for negatives that happen to be very close to the anchor through image-to-text and text-to-image hardness parameters $\beta_1, \beta_2 \geq 0$. Here, $t \to v$ means text-to-image, and $v \to t$ denotes image-to-text.

$$\mathcal{L}^{v\to t} = -\sum_{i=1}^{B} \frac{\mathbf{I}_{p,i}\mathbf{T}^{\top}_{p,i}}{\tau} + \sum_{i=1}^{B} log\left(\sum_{j\neq i} e^{\mathbf{I}_{p,i}\mathbf{T}^{\top}_{p,j}/\tau} \mathcal{W}^{v\to t}_{\mathbf{I}_{p,i}\mathbf{T}_{p,j}}\right) \tag{1}$$

$$\mathcal{L}^{t\to v} = -\sum_{i=1}^{B} \frac{\mathbf{T}_{p,i}\mathbf{I}_{p,i}^{\top}}{\tau} + \sum_{i=1}^{B} log\left(\sum_{j\neq i} e^{\mathbf{T}_{p,i}\mathbf{I}_{p,j}^{\top}/\tau} \mathcal{W}^{t\to v}_{\mathbf{T}_{p,i}\mathbf{I}_{p,j}}\right) \tag{2}$$

$$\mathcal{L}_{DHN-NCE} = \mathcal{L}^{v\to t} + \mathcal{L}^{t\to v} \tag{3}$$

where B is the batch size, τ is the temperature parameter, and the hardness weighting formulas are as follows:

$$\mathcal{W}^{v\to t}_{\mathbf{I}_{p,i}\mathbf{T}_{p,j}} = (B-1) \times \frac{e^{\beta_1 \mathbf{I}_{p,i}\mathbf{T}_{p,j}/\tau}}{\sum_{k\neq i} e^{\beta_1 \mathbf{I}_{p,i}\mathbf{T}_{p,k}/\tau}} \tag{4}$$

$$\mathcal{W}^{t\to v}_{\mathbf{T}_{p,i}\mathbf{I}_{p,j}} = (B-1) \times \frac{e^{\beta_2 \mathbf{T}_{p,i}\mathbf{I}_{p,j}/\tau}}{\sum_{k\neq i} e^{\beta_2 \mathbf{T}_{p,i}\mathbf{I}_{p,k}/\tau}} \tag{5}$$

BiomedCLIP Fine-Tuning. We utilized the public MedPix dataset with different radiological modalities to fine-tune the BiomedCLIP model [33] with DHN-NCE loss. Here, we used the base Vision Transformer and PubMedBERT [33] as the image and text encoders. We cleaned the MedPix dataset by stripping off any special characters, leading and trailing white spaces, and deleting samples with captions of less than 20 characters. All images were resized to 224 × 224 pixels and normalized by the RGB channel means and standard deviations used in the original CLIP model [26]. After performing an 85%:15% split, we ended up with 20,292 training images and 3,515 images for validation. Here, we chose a low learning rate of 1E-6 with a decay rate of 50%, and fine-tuning was done on batches of 64 samples.

2.2 Zero-Shot and Weakly Supervised Medical Image Segmentation

With a fine-tuned BiomedCLIP model, we proposed a zero-shot universal medical image segmentation strategy, which leverages the recent XAI technique, gScoreCAM [6] that provides visual saliency maps of text prompts in corresponding images for CLIP models. While gScoreCAM was shown to outperform gradCAM in natural images in accuracy and specificity, we adopted it in radiological tasks for the first time. Here, for an input image and a text prompt for the target anatomy/pathology, we first obtained an initial, coarse segmentation by post-processing the gScoreCAM map with a conditional random field (CRF) filter [15] in order to produce discrete pixel-wise labels as initial segmentation masks. These generated initial segmentation labels are used to calculate 4 box coordinates (bounding boxes) that enclose each connected contour in the segmentation mask. Finally, we supply these bounding boxes as prompt inputs to SAM in order to produce a pseudo-mask as zero-shot segmentation. In the attempt to further enhance the accuracy of zero-shot segmentation, we used the resulting pseudo-masks as the sole supervision signal to train a Residual UNet [34] in a weakly supervised setting.

2.3 Datasets, Experimental Setup, and Validation Metrics

BiomedCLIP Fine-Tuning Performance. We validated the quality of BiomedCLIP fine-tuning by the accuracy of top 1 and top 2 matching retrievals for both image-to-text and text-to-image directions in the ROCO (Radiology Objects in COntext) dataset [24] which contains ≈ 7,042 multi-modal medical images spanning a myriad of clinical cases. We executed the experiments for 5 runs with a batch size of 50 with shuffling to ensure random bagging of different texts and images within a batch (thus, we get 70,420 shuffled examples). We compared different loss functions for fine-tuning, including the InfoNCE loss [23], DCL [32], HN-NCE [25], and our DHN-NCE loss. For a fair comparison, we trained all the strategies using the same hyperparameters ($\tau = 0.6$, learning rate = 1E-6). For HN-NCE and DHN-NCE, we use the same hardness $\beta_1 = \beta_2 = 0.15$. As baselines, we also included the results of pre-trained BiomedCLIP [33], PMC-CLIP [18], and CLIP [26].

Image Segmentation Accuracy. To validate the zero-shot and weakly supervised segmentation results, as well as different design components of the MedCLIP-SAM framework, we used three public datasets (three different modalities) with segmentation ground truths (segmentation of breast tumor, brain tumor, and lung), which were split for training, validation, and testing. These datasets with their divisions include:

- **Breast Tumor Ultrasound**: Breast Ultrasound Images dataset (BUSI) [1] with 600 benign and malignant tumors images for training only; 65 and 98 images from the UDIAT [5] dataset for validation and testing, respectively.
- **Brain Tumor MRI**: Brain Tumor dataset from [10] consisting of 1,462, 400, and 400 T1-weighted MRIs for training, validation and testing respectively.
- **Lung Chest X-ray**: COVID-19 Radiography Database (COVID-QU-Ex) [11,27] with 16,280, 1,372, and 957 Chest X-ray scans (normal, lung opacity, viral pneumonia, and COVID-19 cases) for training, validation, and testing.

With these datasets, we conducted a detailed comparison of the segmentation quality for the initial labels based on CRF-processed gScoreCAM results, zero-shot pseudo-masks, and weakly supervised results on the aforementioned testing sets. As ablation studies for zero-shot segmentation, we investigated **1)** the impacts of BiomedCLIP fine-tuning and **2)** the choice of gScoreCAM vs. grad-CAM. The ablation studies were performed on the test set of each of the three aforementioned datasets. For a fair comparison, we utilized the same SAM model, target layer, text prompts, and CAM settings of the top 60 channels for all data across different variations. In all experiments, Intersection over Union (IoU), Dice Similarity Coefficient (DSC), and area under the ROC curve (AUC) were used, and paired-sample t-tests were performed to confirm the observations and trends. Here, a p-value < 0.05 indicates a statistically significant difference.

3 Results

3.1 Cross-Modal Retrieval Accuracy and gScoreCAM Vs. GradCAM

The accuracy of cross-modal retrieval (text-to-image and image-to-text) for the ROCO dataset [24] is shown in Table 1 across different losses for fine-tuning BiomedCLIP, with three pre-trained CLIP models as baselines. Paired McNemar statistical tests show that our DHN-NCE significantly outperformed other existing loss functions and pre-trained baseline models (p <0.01). In Table 2, we present the accuracy evaluation for our MedCLIP-SAM zero-shot segmentation with different setups (Pre-trained BiomedCLIP vs. fine-tuned BiomedCLIP and gScoreCAM vs. GradCAM). The comparison demonstrated the great advantages of using gScoreCAM over GradCAM to generate bounding-box prompts for SAM (p <1E-4). Additionally, the benefit of fine-tuning BiomedCLIP with our DHN-NCE loss is further validated with improved segmentation quality across different tasks and image modalities ($p < 0.05$).

Table 1. Top-K cross-modal retrieval accuracy (mean±std) for CLIP models.

Model	Version	*image* → *text* (%)		*text* → *image* (%)	
		Top-1	Top-2	Top-1	Top-2
BiomedCLIP [33]	Pre-trained	81.83 ± 0.20	92.79 ± 0.13	81.36 ± 0.48	92.27 ± 0.14
	InfoNCE [23]	84.21 ± 0.35	94.47 ± 0.19	85.73 ± 0.19	94.99 ± 0.16
	DCL [32]	84.44 ± 0.37	94.68 ± 0.19	85.89 ± 0.16	95.09 ± 0.19
	HN-NCE [25]	84.33 ± 0.35	94.60 ± 0.19	85.80 ± 0.17	95.10 ± 0.19
	DHN-NCE (Ours)	**84.70 ± 0.33**	**94.73 ± 0.16**	**85.99 ± 0.19**	**95.17 ± 0.19**
CLIP [26]	Pre-trained	26.68 ± 0.30	41.80 ± 0.19	26.17 ± 0.20	41.13 ± 0.20
PMC-CLIP [18]	Pre-trained	75.47 ± 0.37	87.46 ± 0.11	76.78 ± 0.11	88.35 ± 0.19

3.2 Zero-Shot and Weakly Supervised Segmentation

In Table 3, we present segmentation accuracy for our proposed method in zero-shot and weakly supervised settings, with fully supervised segmentation as a reference. Note that for zero-shot results, we include a comparison between initial labels generated by gScoreCAM-based saliency maps ("Saliency Maps") and pseudo-masks from SAM ("Saliency Maps + SAM"). Combining BiomedCLIP and SAM demonstrates clear advantages, notably improving segmentation quality for all metrics ($p < 0.05$). Comparing zero-shot results to weakly supervised segmentation, we observe general improvements for X-ray-based lung segmentation. However, the impact on tumor segmentation in breast ultrasound and brain MRI remains unclear, with an AUC boost of ∼2% only for breast ultrasound. While fully supervised DL models currently provide state-of-the-art accuracy for medical image segmentation, our MedCLIP-SAM zero-shot segmentation outperformed ResUNet-based full supervision for breast ultrasound and brain

MRI segmentation. Lung X-ray segmentation, however, showed superior accuracy with the fully supervised method across all metrics. Finally, to provide a qualitative assessment, exemplary segmentation results for zero-shot and weakly supervised settings are shown in Fig. 2 against the original image and ground truths (GTs) across all segmentation tasks.

Table 2. Comparison of zero-shot segmentation accuracy (mean±std) with SAM based on the pre-trained and fine-tuned BiomedCLIP models using gScoreCAM vs. GradCAM techniques for bounding-box generation.

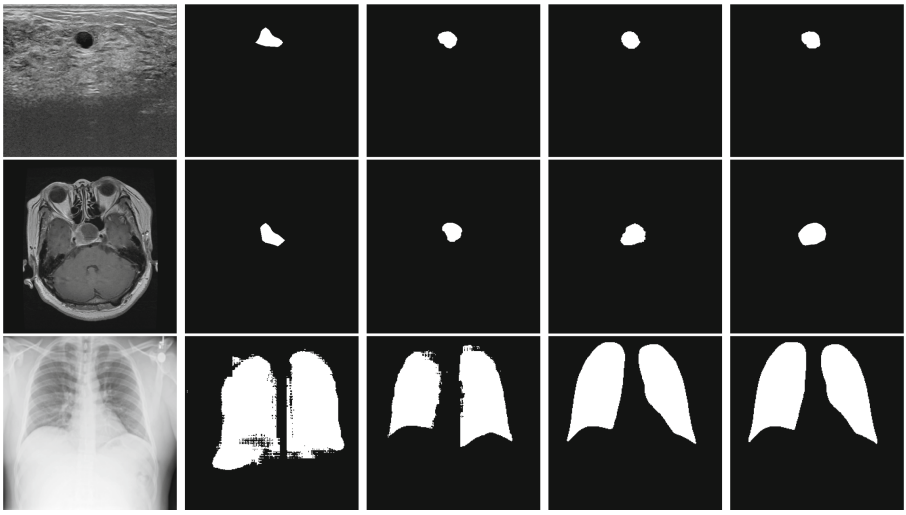
Modality	Model	CAM	IoU (%)	DSC (%)	AUC (%)
Breast Ultrasound	BiomedCLIP	gScoreCAM	56.24 ± 9.25	66.03 ± 8.77	78.59 ± 6.38
		GradCAM	18.16 ± 9.67	23.99 ± 8.24	60.12 ± 6.36
	Ours	**gScoreCAM**	**57.97 ± 8.59**	**67.82 ± 8.26**	**79.31 ± 6.84**
		GradCAM	20.79 ± 9.32	25.65 ± 7.81	62.54 ± 5.22
Brain MRI	BiomedCLIP	gScoreCAM	48.87 ± 6.71	65.13 ± 5.98	79.69 ± 6.12
		GradCAM	26.69 ± 7.45	32.03 ± 5.23	76.04 ± 7.86
	Ours	**gScoreCAM**	**50.30 ± 5.94**	**66.72 ± 5.27**	**81.35 ± 6.33**
		GradCAM	27.07 ± 7.29	33.10 ± 6.91	78.72 ± 7.16
Lung X-ray	BiomedCLIP	gScoreCAM	47.95 ± 10.37	63.21 ± 11.70	77.53 ± 5.49
		GradCAM	22.79 ± 7.35	35.21 ± 10.75	60.19 ± 4.73
	Ours	**gScoreCAM**	**49.06 ± 9.22**	**64.49 ± 9.09**	**78.54 ± 5.64**
		GradCAM	26.45 ± 8.39	39.75 ± 8.44	62.95 ± 5.71

Table 3. Segmentation accuracy (mean±std) for zero-shot and weakly supervised methods against a fully supervised baseline.

Modality	Model	IoU (%)	DSC (%)	AUC (%)
Breast Ultrasound	Saliency Maps	40.43 ± 8.34	51.82 ± 9.60	73.77 ± 7.54
	Saliency Maps + SAM	**57.97** ± 8.59	**67.82 ± 8.26**	79.31 ± 6.84
	Weak supervision-ResUNet [34]	41.68 ± 5.63	58.62 ± 5.66	81.44 ± 4.22
	Full supervision-ResUNet [34]	53.15 ± 8.36	67.29 ± 7.84	**84.74 ± 5.09**
Brain MRI	Saliency Maps	39.12 ± 6.11	53.06 ± 6.34	75.89 ± 6.92
	Saliency Maps + SAM	**50.30 ± 5.94**	**66.72 ± 5.27**	**81.35 ± 6.33**
	Weak supervision-ResUNet [34]	42.17 ± 8.67	58.80 ± 8.63	78.25 ± 5.32
	Full supervision-ResUNet [34]	45.93 ± 7.68	62.57 ± 7.20	79.85 ± 4.87
Lung X-ray	Saliency Maps	35.04 ± 8.40	49.54 ± 9.18	71.94 ± 6.21
	Saliency Maps + SAM	49.06 ± 9.22	64.49 ± 9.09	78.54 ± 5.64
	Weak supervision-ResUNet [34]	76.46 ± 12.03	86.07 ± 8.61	90.76 ± 4.39
	Full supervision-ResUNet [34]	**95.26 ± 4.82**	**97.50 ± 2.84**	**98.38 ± 2.01**

4 Discussion

To the best of our knowledge, our proposed MedCLIP-SAM presents the first framework that integrates CLIP and SAM models toward universal radiological segmentation. By leveraging the latest CAM technique, gScoreCAM, which



(a) Image (b) CLIP-CRF (c) Zero-Shot (d) WSS (e) GT

Fig. 2. Qualitative comparison of segmentation results. CLIP-CRF=CRF processed BiomedCLIP saliency map and WSS=weakly supervised segmentation.

is used in medical imaging for the first time, our method offers a unique solution that allows text-prompt-based interaction, easy adaptation to new data domains/tasks, and data-efficient model (pre-)training. One major contribution of this work lies in the newly devised DHN-NCE loss, which benefits from the synergy of DCL and HN-NCE and has been demonstrated to outperform the state-of-the-art loss functions (see Table 1) to efficiently fine-tune the BiomedCLIP model with a small batch size. Although we only demonstrated its application in unsupervised CLIP model fine-tuning, we will test its application in full model training in the near future. When using BiomedCLIP and gScoreCAM to obtain saliency maps, we used more simplistic keywords for segmentation tasks, such as "brain tumor". However, we also noticed that the quality of these maps could benefit from more sophisticated text prompt engineering, including detailed descriptions (e.g., shape and location of the target anatomy/pathology). This leaves an interesting application of our MedCLIP-SAM framework for interactive radiological education. From the ablation studies, both gScoreCAM and fine-tuned BiomedCLIP positively contributed to the success of our method. Our weakly supervised segmentation only improved the accuracy in X-ray-based lung segmentation. This could be explained by the complex contrast of ultrasound and the 3D nature of the brain MRI, which may be more suitable for 3D segmentation. Notably, the latest MedSAM [22] has demonstrated superior performance for medical applications. However, as it was fine-tuned on large amounts of public medical datasets, which include our test databases, adopting it for our framework will invalidate the "zero-shot" setting. With encouraging results from SAM in our framework, we aim to further explore the incorporation

of MedSAM into MedCLIP-SAM to verify the potential performance enhancement. Finally, we only tested three segmentation tasks and image modalities in this study, and will expand our validation to a broader range of applications and image types.

5 Conclusion

We proposed MedCLIP-SAM, a novel framework that combines CLIP and SAM foundation models to obtain text-prompt-based universal medical image segmentation. The interactive nature of the method provides a unique venue to allow human interaction. In addition, our newly proposed DHN-NCE loss could potentially benefit broader applications. Our comprehensive experiments demonstrated excellent performance of the proposed framework, which possesses great potential for clinical adoption upon future improvements.

Acknowledgments. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

Disclosure of Interests. The authors have no competing interests.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in Brief **28**, 104863 (2020). https://doi.org/10.1016/j.dib.2019.104863
2. Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M., Adebayo, J., Li, M.D., Kalpathy-Cramer, J.: Assessing the (un)trustworthiness of saliency maps for localizing abnormalities in medical imaging (2021)
3. Bae, W., Noh, J., Kim, G.: Rethinking class activation mapping for weakly supervised object localization. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16. pp. 618–634. Springer (2020)
4. Baevski, A., Babu, A., Hsu, W.N., Auli, M.: Efficient self-supervised learning with contextualized target representations for vision, speech and language. In: International Conference on Machine Learning. pp. 1416–1429. PMLR (2023)
5. Byra, M., Jarosik, P., Szubert, A., Galperin, M., Ojeda-Fournier, H., Olson, L., O'Boyle, M., Comstock, C., Andre, M.: Breast mass segmentation in ultrasound with selective kernel U-Net convolutional neural network. Biomed Signal Process Control **61** (Jun 2020)
6. Chen, P., Li, Q., Biaz, S., Bui, T., Nguyen, A.: gscorecam: What is clip looking at? In: Proceedings of the Asian Conference on Computer Vision (ACCV) (2022)
7. Chen, T., Mai, Z., Li, R., lun Chao, W.: Segment anything model (sam) enhanced pseudo labels for weakly supervised semantic segmentation (2023)
8. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. Advances in neural information processing systems **33**, 22243–22255 (2020)

9. Chen, Z., Xu, Q., Liu, X., Yuan, Y.: Un-sam: Universal prompt-free segmentation for generalized nuclei images (2024)
10. Cheng, J.: brain tumor dataset (4 2017). https://doi.org/10.6084/m9.figshare.1512427.v5
11. Chowdhury, M.E.H., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M.A., Mahbub, Z.B., Islam, K.R., Khan, M.S., Iqbal, A., Emadi, N.A., Reaz, M.B.I., Islam, M.T.: Can ai help in screening viral and covid-19 pneumonia? IEEE Access **8**, 132665–132676 (2020). https://doi.org/10.1109/ACCESS.2020.3010287
12. Hu, X., Xu, X., Shi, Y.: How to efficiently adapt large segmentation model(sam) to medical images (2023)
13. Huang, Z., Liu, H., Zhang, H., Li, X., Liu, H., Xing, F., Laine, A., Angelini, E., Hendon, C., Gan, Y.: Push the boundary of sam: A pseudo-label correction framework for medical segmentation (2023)
14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything (2023)
15. Kraehenbuehl, P., Koltun, V.: Parameter learning and convergent inference for dense random fields. In: Dasgupta, S., McAllester, D. (eds.) Proceedings of the 30th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 28, pp. 513–521. PMLR, Atlanta, Georgia, USA (17–19 Jun 2013)
16. Li, S., Cao, J., Ye, P., Ding, Y., Tu, C., Chen, T.: Clipsam: Clip and sam collaboration for zero-shot anomaly segmentation (2024)
17. Li, Y., Wang, H., Duan, Y., Li, X.: Clip surgery for better explainability with enhancement in open-vocabulary tasks (2023)
18. Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., Xie, W.: Pmc-clip: Contrastive language-image pre-training using biomedical documents (2023)
19. Liu, J., Lin, Z., Padhy, S., Tran, D., Bedrax Weiss, T., Lakshminarayanan, B.: Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. Advances in Neural Information Processing Systems **33**, 7498–7512 (2020)
20. Liu, X., Huang, X.: Weakly supervised salient object detection via bounding-box annotation and sam model. Electronic Research Archive **32**(3), 1624–1645 (2024)
21. Loquercio, A., Segu, M., Scaramuzza, D.: A general framework for uncertainty estimation in deep learning. IEEE Robotics and Automation Letters **5**(2), 3153–3160 (2020)
22. Ma, J., Wang, B.: Segment anything in medical images. ArXiv **abs/2304.12306** (2023)
23. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
24. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.: Radiology objects in context (roco): A multimodal image dataset. In: CVII-STENT/LABELS@MICCAI (2018)
25. Radenovic, F., Dubey, A., Kadian, A., Mihaylov, T., Vandenhende, S., Patel, Y., Wen, Y., Ramanathan, V., Mahajan, D.: Filtering, distillation, and hard negatives for vision-language pre-training. arXiv:2301.02280 (2023)
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021)
27. Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Abul Kashem, S.B., Islam, M.T., Al Maadeed, S., Zughaier, S.M., Khan, M.S., Chowdhury,

M.E.: Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. Computers in Biology and Medicine **132**, 104319 (2021). https://doi.org/10.1016/j.compbiomed.2021.104319
28. Robinson, J., Chuang, C.Y., Sra, S., Jegelka, S.: Contrastive learning with hard negative samples (2021)
29. Siuly, S., Zhang, Y.: Medical big data: neurological diseases diagnosis through medical data analysis. Data Science and Engineering **1**, 54–64 (2016)
30. Taleb, A., Lippert, C., Klein, T., Nabi, M.: Multimodal self-supervised learning for medical image analysis. In: International conference on information processing in medical imaging. pp. 661–673. Springer (2021)
31. Yang, X., Gong, X.: Foundation model assisted weakly supervised semantic segmentation (2023)
32. Yeh, C.H., Hong, C.Y., Hsu, Y.C., Liu, T.L., Chen, Y., LeCun, Y.: Decoupled contrastive learning (2022)
33. Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Lungren, M.P., Naumann, T., Poon, H.: Large-scale domain-specific pretraining for biomedical vision-language processing (2023)
34. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual u-net. IEEE Geoscience and Remote Sensing Letters **15**(5), 749-753 (May 2018). https://doi.org/10.1109/lgrs.2018.2802944