Lite Class-prompt Tiny-VIT for Multi-Modality Medical Image Segmentation

Haotian Guan[1†], Bingze Dai[1†][0000−0003−2199−6006], and Jiajing Zhang[1][0000−0002−4981−3534]

Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, China
jiajingz@connect.hku.hk

Abstract. The increasing demand for accurate medical image segmentation is crucial for alleviating the workload of doctors and enhancing diagnostic accuracy, particularly in low-income countries with limited computational resources. This study investigates the application of a novel deep learning model, class-prompt Tiny-VIT, to segment various medical image modalities using a laptop. The primary focus is on the challenges posed by the significant differences across image modalities, which render a unified model ineffective in handling certain modalities like positron emission tomography (PET) with high dice similarity on the segmentation task. Experimental results demonstrate that the class prompt, a simplified yet efficient method, can effectively boost model performance on modalities such as PET and microscopy, achieving improved overall segmentation accuracy. This research holds significant potential for the practical implementation of medical image segmentation in resource-constrained settings, and underlines the importance of developing deep learning algorithms tailored to specific medical imaging modalities.

Keywords: Multi-modality · Medical image segmentation · Class prompt · TinyVIT .

1 Introduction

Medical image segmentation plays a crucial role in computer-aided diagnosis, treatment planning, disease progression monitoring, image-guided interventions, and personalized medicine. The accurate delineation of anatomical structures and pathological regions is essential for effective clinical decision-making [1].

Various deep-learning based semantic segmentation models have been proposed [2, 3] while most existing fundamental segmentation models are mainly based on natural images. Various machine learning based models were proposed to cope with different specific segmentation tasks including brain, liver, tumour, cell, lung, cardiac, vascular etc [4–11]with different imaging modalities such as MRI, OCT, ultrasound, X-Ray, ultrasound etc and have demonstrated remarkable success in medical image segmentation tasks which helps on various medical

tasks including tumor diagnostic [12, 13], vessel and tissue characterization and so on [14–17].

However, the complexity and diversity of medical imaging modalities, along with the inherent variability in anatomical structures, make it difficult to design a robust and efficient segmentation model that can perform well across different imaging scenarios [18]. Instead of focusing on specialized models for single tasks, researchers have explored more generalized models that can manage multiple scenarios like SAM [19], this trend also lead to the development of generalized methods for medical image segmentation across different modalities with U-Net architecture [18, 20] and models inspired by SAM like MedSAM [21–23] and so on. While in the meantime, these methods often require substantial computational resources, which may not be feasible for deployment on resource-constrained devices such as laptops. To enable real-time processing and edge-machine use with such machine learning models, the size and computation complexity need to be reduced. To this end, MobileSAM [24], and EfficientViT-SAM [25], TinyVIT [26] have shown promising results in terms of both accuracy and computational efficiency. These methods employ self-attention mechanisms to capture global and local contextual information to improve performance with comparable small models. However, despite their success, these methods still face limitations in terms of model size, computational complexity, and adaptability to different imaging modalities, particularly when considering the deployment of these methods on resource-constrained devices such as laptops with an 8G CPU without GPU that is commonly used in clinical.

To solve the computation cost and multi-modality generalization challenges mentioned above, inspired by TinyVIT [26] and Vision Transformer (ViT) architecture [27] which have demonstrated great potential in computer vision tasks, here we propose a novel approach Class-prompted TinyVIT network for medical image segmentation across different imaging modalities, with high accuracy and low computation cost that is capable of running on an 8G CPU device in almost real-time. The Class-prompted TinyVIT is inspired by the actual divisions in hospitals where different modalities are assigned to specific doctors instead of letting the same doctor to read all modalities images. By this class-prompt method, the model would acquire the modality class information while using the same model with tuned parameters with specific modality while keep the similar model size.

By adapting this architecture, we aim to provide a compact and efficient solution that can be deployed on devices with limited computational resources. Our primary contribution lies in introducing class-prompt to the lite TinyVIT model to cope with the multi-modality medical image segmentation across different imaging modalities effectively as specialized models. The combination of compact model size, high accuracy, and compatibility with resource-constrained CPUs makes our approach a promising solution for real-world medical imaging applications. Furthermore, we conduct experiments on the provided dataset to validate the effectiveness of our approach. The results demonstrate superior segmentation accuracy and computational efficiency, highlighting the potential

of our approach in advancing medical image analysis for edge-device application. In summary, this paper presents a novel class-prompted TinyVIT-based approach for medical image segmentation that addresses the challenges of model size, computational complexity, and adaptability to diverse imaging modalities. Our solution holds the potential to significantly impact the field of medical image analysis, particularly in the context of deployment on resource-constrained devices.

2 Method

We introduce a pioneering class-prompt-based methodology aimed at enhancing segmentation efficacy across a spectrum of medical imaging modalities. Drawing inspiration from the specialized organizational structure of hospitals, where domain experts are assigned to interpret distinct imaging modalities, our objective is to cultivate expert models tailored to each modality. However, the deployment of 11 individual models on a laptop proves unfeasible due to memory constraints. To address this challenge, we propose a prompt-driven approach that effectively communicates the current input modality to the model, thereby transforming it into an adept specialist for the specific modality under consideration. This innovative strategy enables us to uphold a concise and efficient model architecture while attaining superior segmentation performance across diverse medical imaging tasks.

2.1 Preprocessing

The original dataset consists of eleven modalities with unbalanced data samples as illustrated in Fig. 1. Among those, CT holds the biggest portion with 1218411 items. The size of the whole training dataset is about 6TB with 1,000,000+ image-mask pairs, covering 10 medical image modalities and more than 20 cancer types. To deal with such a large and imbalanced dataset, to avoid redundant and long training time costs, we first sampled the original dataset. Though we can try to train on the whole dataset to get a more comprehensive model, the time and electricity cost for the training process would be very burdensome. As a result, not only to make the model lite, we decide to make the training process also lite so that researchers with a single normal GPU can train it within reasonable time and calculation cost. Here we first sampled the original large dataset less than 1/10 of its original size. As CT images are far more than the number of other modalities, we randomly sampled CT images with 1/50 and randomly sampled other modalities to 1/10 of original numbers as shown in Figure 1. After sampling, the sampled dataset size drops to less than 300 GB.

Not only the number of each modality is different, the image and mask size in different modalities are also different which draws a problem to input the same model for training. As a result, to fit the training model, we first resize all the input images to 256*256 pixels to keep consistence so that we can use the same lite model. Boxes are generated using ground truth. Ground truth are covered completely by the box. In theory, area outside the box should not be segmented.

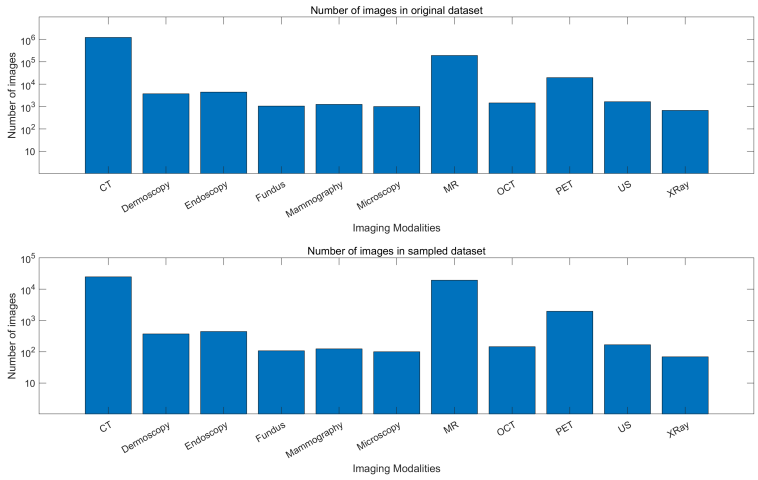

Fig. 1. Number of images of different modalities in original dataset and sampled dataset for training.

2.2 Proposed Method

Class Prompt: Various image classifier models have been explored, including traditional networks, deep learning, transfer learning, self-supervised learning, and more [28–31]. However, most of these models use large architectures, making them unsuitable for tasks with limited computational resources.

In our model, we address this limitation by incorporating a classifier that leverages the TinyVIT encoder. This approach notably diminishes both the model's size and parameter count. To enhance the classifier's efficacy further, we have integrated a three-layer multilayer perceptron (MLP) network as the modality classifier head. This augmentation exploits the insights derived from the encoder structure. As a result, our model showcases outstanding performance.

Specifically, we equipped TinyVIT with a class prediction head, which is implemented as a multilayer perceptron (MLP) that starts with an input dimension of 256. This MLP features a hidden layer dimension of 256, and is designed to classify data into one of 11 distinct categories. The architecture includes three layers, to process and refine the information through successive transformations. The first layer takes the input vector $\mathbf{x}$ of dimension 256 and transforms it to a hidden state $\mathbf{h}_1$ using a linear transformation followed by a non-linear activation function (ReLU). The MLP is represented as:

$$\mathbf{h}_1 = ReLU(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) \tag{1}$$

$$\mathbf{h}_2 = ReLU(\mathbf{W}_2\mathbf{h}_1 + \mathbf{b}_2) \tag{2}$$

$$\mathbf{y} = \mathbf{W}_3\mathbf{h}_2 + \mathbf{b}_3 \tag{3}$$

where $\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3$ are the weight matrixes and $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ is the bias vectors. This structure allows the classifier to effectively learn and make predictions by capturing complex patterns and relationships in the data, making it a vital component in achieving high accuracy in our classification tasks.

Class-prompt Tiny-VIT Class prompt takes the categories from the classification head in encoder and returns a prompt for mask decoder. The pretrained encoder makes sure that images modalities can be correctly identified. The small size of class prompt encoder is specially designed for running inference on laptop. The prompt is added to the box prompt as the input of decoder. Our decoder is a transformer with three blocks.

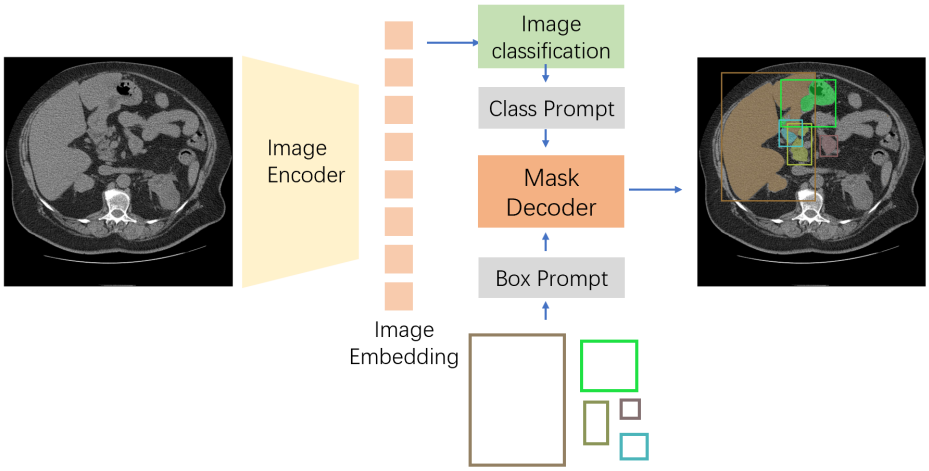

Fig. 2. Class-prompted TinyVIT Network architecture. Class prompt takes the categories from the classification head in encoder and returns a prompt for mask decoder.

Loss function: We use the summation between Dice loss and focal loss because compound loss functions have been proven to be robust in various medical image segmentation tasks [32].

a) Dice loss is a performance metric derived from the Dice coefficient, which is commonly used to gauge the similarity between two samples. Specifically tailored for the field of medical image segmentation, the Dice loss function is particularly effective in handling class imbalance, a frequent challenge where the region of interest occupies a significantly smaller portion of the image compared to the background. The Dice loss is calculated as

$$L_{Dice} = 1 - \frac{2 \times |X \cap Y|}{|X| + |Y|} \tag{4}$$

where X and Y represent the binary prediction and ground truth masks, respectively. This loss function ensures that the model is not only predicting the classes accurately but also aligning closely with the actual contours and boundaries of the regions of interest in the images.

b) Focal loss is an advanced adaptation of the cross-entropy loss, specifically designed to address the prevalent issue of class imbalance by concentrating more on difficult, misclassified examples. This is particularly useful in image segmentation tasks where there is a significant imbalance between different classes. The focal loss function is mathematically represented as:

$$L_{Focal} = -\alpha_t(1 - p_t)^\gamma \log(p_t) \tag{5}$$

Here, p_t is the probability that the model assigns to the ground truth class. The parameter γ is the focusing parameter, which scales how much the function focuses on hard examples. The term $(1-p_t)^\gamma$ decreases the loss contribution from easy examples and increases the importance of correcting misclassified examples. α_t is a balancing factor that can be used to give more focus to rare classes. This formulation helps in fine-tuning the model's predictions, ensuring that it not only achieves high accuracy but also improves performance on the more challenging aspects of the segmentation task.

From the perspective of inference efficiency, our simple MLP structure can be easily implemented on CPU-only machines without complex acceleration strategies.

2.3 Post-processing

The size of the mask outputted from the model is unified to 256 by 256. In order to obtain the mask for the original medical image, we resize the outputted mask to the original size with bilinear interpolation. The model will produce a prediction for each box with its index. To save the complete result, all boxes and corresponding segmentation are saved as an overlay.

3 Experiments

3.1 Dataset and evaluation measures

We used the challenge dataset for model development, including 11 modalities and both 2D and 3D images.

The evaluation metrics include two accuracy measures—Dice Similarity Coefficient (DSC) and Normalized Surface Dice (NSD)—alongside one efficiency measure—running time.The Dice Similarity Coefficient (DSC) is formulated as follows:

$$DSC = \frac{2 \times |P \cap G|}{|P| + |G|}$$

where P represents the set of pixels in the predicted segmentation mask, G represents the set of pixels in the ground truth segmentation mask, $|\cdot|$ denotes the cardinality or size of the set, $\cap$ denotes the intersection of sets.

The formula for Normalized Surface Dice (NSD) is given by:

$$NSD = 1 - \frac{2 \times Surface(A \cap B)}{Surface(A) + Surface(B)}$$

where A and B are the segmentation mask and ground truth being compared. $Surface(A)$ and $Surface(B)$ represent the surface areas of masks A and B respectively. $Surface(A \cap B)$ represents the surface area of the intersection of masks A and B.

This formula calculates the Dice similarity coefficient between the surfaces of the two masks and normalizes it by the average surface area of the two masks. A higher NSD value indicates a better overlap between the surfaces of the two masks.

3.2 Implementation details

Environment settings The development environments and requirements are presented in Table 1.

Table 1. Development environments and requirements. (mandatory table)

System	Ubuntu 18.04.5 LTS
CPU	Intel Xeon W-2225 10C/20T 4.1Ghz
RAM	32GB DDR4-3200 ECC RDIMM
GPU (number and type)	One NVIDIA GeForce RTX 3090 24G
CUDA version	10.496
Programming language	Python 3.80
Deep learning framework	torch 2.0, torchvision 0.2.2
Specific dependencies	None
Code	None

Training protocols 1. Data augmentation: in the domain of multimodal medical image segmentation, the strategic implementation of data augmentation stands as a pivotal factor in augmenting model generalization and precision. By adhering to a data augmentation rate of 0.5, incorporating stochastic horizontal flips (fliplr) and vertical flips (flipud), we enrich the training dataset by introducing nuanced variations within the multimodal input images and their corresponding segmentation maps. These meticulous operations empower the model to discern and delineate anatomical structures resilient to horizontal and vertical transformations, thereby fortifying its resilience and efficacy across a spectrum of modalities. Through these refined data augmentation methodologies, our primary objective is to enhance the model's proficiency in accurately segmenting intricate anatomical entities in multimodal medical images, thereby advancing its efficacy within clinical frameworks.

2. Data sampling strategy: during the model training phase, the challenge organizer's publicly available training data was partitioned randomly into training and validation sets in a 7:3 ratio. Specifically, for the 2D image modality, random sampling was directly employed. However, in the case of 3D image modalities such as CT, MRI, and PET, the conventional approach of preprocessing involves segmenting each 3D image into multiple consecutive 2D slices, leading to a proliferation of redundant data due to the significant similarity between adjacent slices. To address this issue, a uniform interval sampling technique was implemented to extract 2D slices from the 3D images, with a fixed spacing of 5 between neighboring samples. This method effectively reduces data redundancy and enhances training efficiency by minimizing the inclusion of highly similar information found in adjacent slices.

3. Optimal model selection criteria: optimal model selection criteria are particularly critical when tasked with segmenting multimodal medical images. In this context, the criteria must be tailored to the nuances of medical image analysis. Given the complexity and variability of medical data, the selected criteria should prioritize robustness and generalizability across different imaging modalities, such as MRI, CT, and PET scans. Here we used Dice similarity coefficient, Jaccard index, and Normalized Surface Distance(NSD) to evaluate segmentation performance in medical imaging tasks. Besides, inferences time is also included in the selection program, giving the same weight as the segmentation performance.

Table 2. Training protocols. (mandatory table)

Pre-trained Model	MedSAM [21]
Batch size	64
Patch size	256×256×3
Total epochs	20
Optimizer	AdamW [33]
Initial learning rate (lr)	5e-4
Lr decay schedule	ReduceLROnPlateau
Training time	75 hours
Loss function	Dice Loss, Cross Entropy Loss, Focal Loss
Number of model parameters	10.99M
Number of flops	2.2G
CO_2eq	15 Kg

4 Results and discussion

4.1 Quantitative results on validation set

Table 4 explains the efficiency of all models by computing the inference time on 14 representative images across all image modalities. On laptop with CPU, 3D

images take much longer inference time since the model deals with segmentation frame by frame. The bottleneck is also on resizing large images to regular input size 256 by 256. Overall, baseline model runs fastest and ablation study is comparable in running time. Our proposed method runs 1.3% to 16.7% slower than baseline model.

Table 3. Quantitative evaluation results.

Target	Baseline		Ablation Study		Proposed	
	DSC(%)	NSD(%)	DSC(%)	NSD(%)	DSC(%)	NSD (%)
CT	89.1	91.03	89.69	91.29	92.26	94.90
MR	83.28	86.1	81.39	83.42	89.63	93.37
PET	55.1	29.12	63.19	46.04	70.28	56.88
US	94.78	96.81	92.92	96.3	94.77	96.81
X-Ray	75.83	80.39	76.7	82.53	76.74	82.53
Dermotology	92.47	93.85	92.45	94.01	93.73	94.01
Endoscopy	96.04	98.11	95.41	97.56	96.04	98.11
Fundus	94.8	96.41	94.56	96.2	94.81	96.41
Microscopy	61.63	65.39	73.76	79.34	73.76	79.34
Average	82.56	81.91	84.45	85.18	86.89	88.04

The results of our quantitative experiments are shown in Table 3. The table presents a comprehensive quantitative evaluation of three models: baseline, ablation model, and our class-prompt Tiny-VIT. Each model's performance is assessed based on Dice Similarity Coefficient (DSC) and Normalized Surface Dice (NSD) metrics across various imaging modalities including CT, MR, PET, US, X-ray, dermatology, endoscopy, fundus, and microscopy.

Upon the quantitative results, it is evident that our class-prompt Tiny-VIT consistently outperforms both the baseline and ablation model across most modalities. Specifically, compared with the baseline, our class-prompt Tiny-VIT demonstrates superior DSC of 3.54%, 7.62%, 27.55%, 1.20%, 1.36%, and 19.68% for CT, MR, PET, X-ray, dermatology, and microscopy, respectively. compared with the baseline, our class-prompt Tiny-VIT demonstrates superior NSD of 4.25%, 8.44%, 95.32%, 2.66%, 0.17%, and 21.33% for CT, MR, PET, X-ray, dermatology, and microscopy, respectively. The improvement indicates our efficacy in accurately segmenting, especially for PET and microscopy images, which is also evidenced by the normalized radargram. As for ultrasound, endoscopy, and fundus images, our class-prompt Tiny-VIT can reach segmentation performance in line with the baseline, thus demonstrating that our improvement in the majority of modalities does not compromise the segmentation performance of the minority modalities. Notably, our class-prompt Tiny-VIT achieves the highest average DSC and NSD values of 86.89% and 88.04% respectively, showcasing its overall effectiveness in comparison to the baseline and ablation modal.

To further demonstrate the role of class prompts, we conducted three fine-tuning experiments. We fine-tuned the baseline on PET, ultrasound, and X-ray

datasets separately to simulate the effect when the model focuses only on a certain class of modes. The DSC and NSD radargrams of the fine-tuning results are shown in Fig. 3. The results show that continual fine-tuning on the ultrasound and X-ray datasets alone does not improve the model performance, but rather leads to a degradation of it. Therefore, it is not feasible to simply fine-tune on each class and then integrate the fine-tuned models of each class. However, the original baseline performs poorly on certain modalities (e.g., PET), and further fine-tuning on these modalities can significantly boost the segmentation performance on that modality, which is not achievable with multimodal mixed training. Therefore, it is worthwhile to enable the model to have an independent perception of each class and to make class-specific processing.

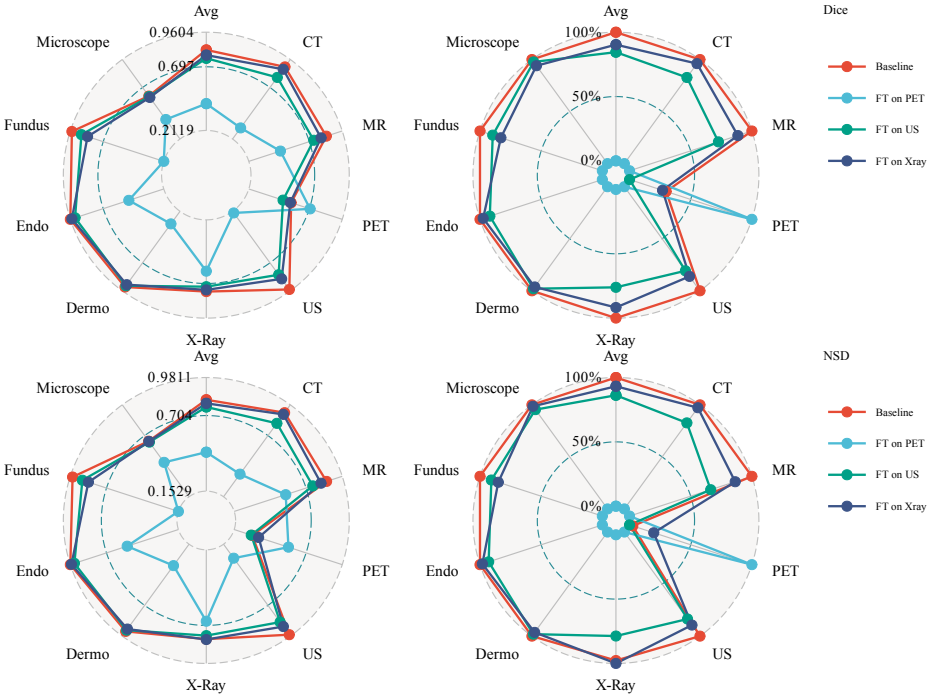

Fig. 3. The Dice result (the first row) and NSD result (the second row) of the baseline model, and baseline model fine-tuned (FT) on PET, Ultrasound, and X-ray datasets, respectively. The left plot shows the raw numerical comparison and the right plot shows the normalized comparison, where the maximum Dice or NSD on each modality is scaled as 1 and the minimum Dice or NSD on each modality is scaled as 0.

Table 4. Quantitative evaluation of segmentation efficiency in terms of running time (s).

Case ID	Size	Num. Objects	Baseline	Ablation Study	Proposed
3DBox_CT_0566	(287, 512, 512)	6	373.6	373.5	381.2
3DBox_CT_0888	(237, 512, 512)	6	92.4	92.3	95.1
3DBox_CT_0860	(246, 512, 512)	1	10.3	10.1	12.7
3DBox_MR_0621	(115, 400, 400)	6	143.6	143.2	150.5
3DBox_MR_0121	(64, 290, 320)	6	91.3	91.5	95.3
3DBox_MR_0179	(84, 512, 512)	1	10.5	10.7	12.2
3DBox_PET_0001	(264, 200, 200)	1	5.8	5.9	7.4
2DBox_US_0525	(256, 256, 3)	1	0.6	0.6	0.7
2DBox_X-Ray_0053	(320, 640, 3)	34	1.9	2.0	2.1
2DBox_Dermoscopy_0003	(3024, 4032, 3)	1	0.8	0.9	1.1
2DBox_Endoscopy_0086	(480, 560, 3)	1	0.6	0.6	0.7
2DBox_Fundus_0003	(2048, 2048, 3)	1	0.8	0.8	0.8
2DBox_Microscope_0008	(1536, 2040, 3)	19	1.7	1.6	1.8
2DBox_Microscope_0016	(1920, 2560, 3)	241	12.9	13.5	14.1

4.2 Qualitative results on validation set

Here we show some examples with good segmentation results and two examples with bad segmentation results. The good cases can almost perfectly segment the region of interest with high DSC ad NSD. Fig. 4 shows the examples with good segmentation results.

Fig. 5 shows the examples with bad segmentation results. For the bad performance cases,some of the segmentation masks cannot fully cover the correct area and some segmentation masks covers more than the region of interest.

4.3 Segmentation efficiency results on validation set

Here we compare the segmentation efficiency based on the time cost on the Codabench platform. On the validation set, the proposed method cost 03:08 min to segment all images and the baseline model costs 03:35 min on validation set. For single cases, please refer to Table.4.

4.4 Results on final testing set

This is a placeholder. We will announce the testing results during CVPR (6.17-18)

4.5 Limitation and future work

In this work, we want to show that class prompt helps with the segmentation of medical images from the perspective of doctors. Unlike natural images, medical images are obtained based on physics. A universal model that understands how

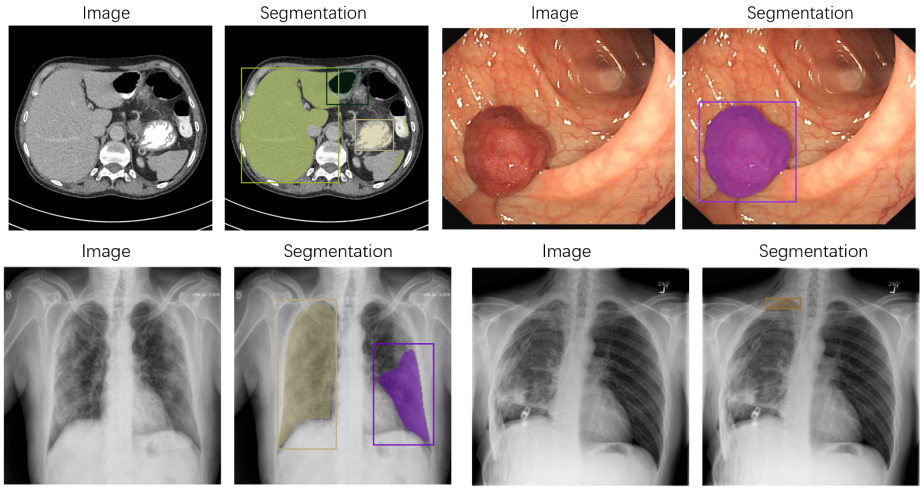

Fig. 4. Examples with good segmentation results

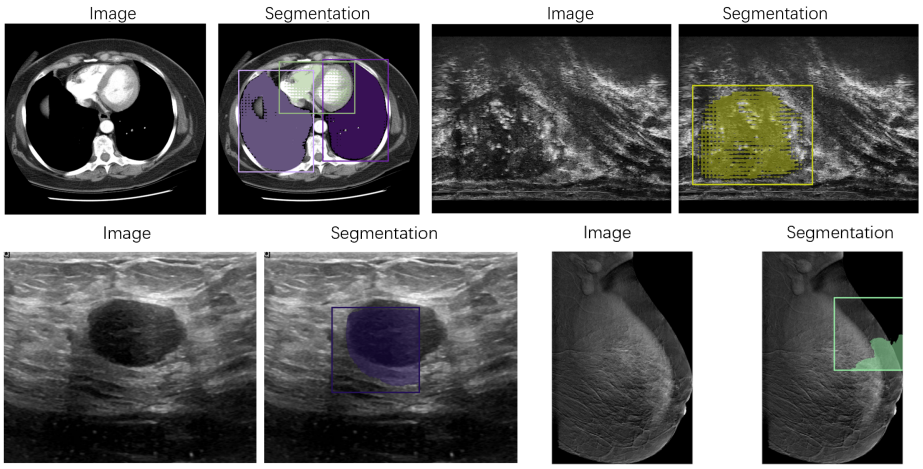

Fig. 5. Examples with bad segmentation results

all modalities work is difficult to achieve. Prompts such as image modality are crucial as patients need to be divided to different departments in a hospital. The class prompt is now built with a 3 layer MLP module. We will design more complicated and reasonable prompt that better suits medical images. Currently the proposed class-prompt TinyVIT is a structure based on TinyVIT. The transformer architecture is still a heavy burden for segmenting medical images on laptop. A smaller model with the same or better precision is highly demanded.

5 Conclusion

Segmenting medical images using laptop is indispensable for alleviating the workload of doctors and improving diagnosing accuracy especially in low-income countries. The main findings and results show that differences across image modalities are huge and a unified model cannot handle modalities such as PET with high dice similarity on segmentation task. Class prompt, as a simple network, can efficiently boost model performance on PET and thus leading to better accuracy overall.

Acknowledgements We thank all the data owners for making the medical images publicly available and CodaLab [34] for hosting the challenge platform.

References

1. G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, p. 60–88, Dec. 2017. [Online]. Available: http://dx.doi.org/10.1016/j.media.2017.07.005 1
2. A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," in *Proceedings of the International Conference on Computer Vision*, 2023, pp. 4015–4026. 1
3. Y. Mo, Y. Wu, X. Yang, F. Liu, and Y. Liao, "Review the state-of-the-art technologies of semantic segmentation based on deep learning," *Neurocomputing*, vol. 493, pp. 626–646, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231222000054 1
4. R. Vivanti, A. Ephrat, L. Joskowicz, O. Karaaslan, N. Lev-Cohain, and J. Sosna, "Automatic liver tumor segmentation in follow-up ct studies using convolutional neural networks," in *Proc. patch-based methods in medical image processing workshop*, vol. 2, 2015, p. 2. 1
5. W. Li, F. Jia, and Q. Hu, "Automatic segmentation of liver tumor in ct images with deep convolutional neural networks," *Journal of Computer and Communications*, vol. 3, no. 11, pp. 146–151, 2015. 1
6. F. Wu and X. Zhuang, "Cf distance: a new domain discrepancy metric and application to explicit domain adaptation for cross-modality cardiac image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 4274–4285, 2020. 1

7. Y. Onishi, A. Teramoto, M. Tsujimoto, T. Tsukamoto, K. Saito, H. Toyama, K. Imaizumi, and H. Fujita, "Multiplanar analysis for pulmonary nodule classification in ct images using deep convolutional neural network and generative adversarial networks," *International journal of computer assisted radiology and surgery*, vol. 15, pp. 173–178, 2020. 1
8. T.-H. Song, V. Sanchez, H. EIDaly, and N. M. Rajpoot, "Dual-channel active contour model for megakaryocytic cell segmentation in bone marrow trephine histology images," *IEEE transactions on biomedical engineering*, vol. 64, no. 12, pp. 2913–2923, 2017. 1
9. V. Cherukuri, P. Ssenyonga, B. C. Warf, A. V. Kulkarni, V. Monga, and S. J. Schiff, "Learning based segmentation of ct brain images: application to postoperative hydrocephalic scans," *IEEE transactions on biomedical engineering*, vol. 65, no. 8, pp. 1871–1884, 2017. 1
10. A. Farshad, Y. Yeganeh, P. Gehlbach, and N. Navab, "Y-net: A spatiospectral dual-encoder networkfor medical image segmentation," 2022. 1
11. S. Diao, J. Su, C. Yang, W. Zhu, D. Xiang, X. Chen, Q. Peng, and F. Shi, "Classification and segmentation of oct images for age-related macular degeneration based on dual guidance networks," *Biomedical Signal Processing and Control*, vol. 84, p. 104810, 07 2023. 1
12. B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014. 2
13. S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in mri images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1240–1251, 2016. 2
14. E. Gibson, F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. Davidson, S. P. Pereira, M. J. Clarkson, and D. C. Barratt, "Automatic multi-organ segmentation on abdominal ct with dense v-networks," *IEEE transactions on medical imaging*, vol. 37, no. 8, pp. 1822–1834, 2018. 2
15. Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer, 2016, pp. 424–432. 2
16. H. Guan, J. Dong, and W.-N. Lee, "Towards real-time training of physics-informed neural networks: Applications in ultrafast ultrasound blood flow imaging," *arXiv preprint arXiv:2309.04755*, 2023. 2
17. B. Dai, "Power doppler ultrasound for peripheral perfusion imaging," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2023. 2
18. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015. 2
19. A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023. 2
20. R. Azad, E. K. Aghdam, A. Rauland, Y. Jia, A. H. Avval, A. Bozorgpour, S. Karimijafarbigloo, J. P. Cohen, E. Adeli, and D. Merhof, "Medical image segmentation review: The success of u-net," 2022. 2
21. J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024. 2, 8

22. M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, "Segment anything model for medical image analysis: an experimental study," *Medical Image Analysis*, vol. 89, p. 102918, 2023. 2
23. J. Wu, W. Ji, Y. Liu, H. Fu, M. Xu, Y. Xu, and Y. Jin, "Medical sam adapter: Adapting segment anything model for medical image segmentation," 2023. 2
24. C. Zhang, D. Han, Y. Qiao, J. U. Kim, S.-H. Bae, S. Lee, and C. S. Hong, "Faster segment anything: Towards lightweight sam for mobile applications," *arXiv preprint arXiv:2306.14289*, 2023. 2
25. Z. Zhang, H. Cai, and S. Han, "Efficientvit-sam: Accelerated segment anything model without performance loss," in *CVPR Workshop: Efficient Large Vision Models*, 2024. 2
26. K. Wu, J. Zhang, H. Peng, M. Liu, B. Xiao, J. Fu, and L. Yuan, "Tinyvit: Fast pretraining distillation for small vision transformers," 2022. 2
27. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. 2
28. L. Cai, J. Gao, and D. Zhao, "A review of the application of deep learning in medical image classification and segmentation," *Annals of translational medicine*, vol. 8, no. 11, 2020. 4
29. B. Dai, T. Qiu, and K. Ye, "Foliar disease classification," 2020. 4
30. S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen *et al.*, "Big self-supervised models advance medical image classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3478–3488. 4
31. H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: a literature review," *BMC medical imaging*, vol. 22, no. 1, p. 69, 2022. 4
32. J. Ma, J. Chen, M. Ng, R. Huang, Y. Li, C. Li, X. Yang, and A. L. Martel, "Loss odyssey in medical image segmentation," *Medical Image Analysis*, vol. 71, p. 102035, 2021. 5
33. I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017. 8
34. Z. Xu, S. Escalera, A. Pavão, M. Richard, W.-W. Tu, Q. Yao, H. Zhao, and I. Guyon, "Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform," *Patterns*, vol. 3, no. 7, p. 100543, 2022. 13

Table 5. Checklist Table. Please fill out this checklist table in the answer column.

Requirements	Answer
A meaningful title	Yes
The number of authors (≤6)	3
Author affiliations and ORCID	Yes/No
Corresponding author email is presented	Yes
Validation scores are presented in the abstract	Yes
Introduction includes at least three parts: background, related work, and motivation	Yes
A pipeline/network figure is provided	Figure 2
Pre-processing	Page 3,4
Strategies to data augmentation	Page 7
Strategies to improve model inference	Page 8
Post-processing	Page 6
Environment setting table is provided	Table 1
Training protocol table is provided	Table 2
Ablation study	Page 9
Efficiency evaluation results are provided	Table 4
Visualized segmentation example is provided	Figure 4,5
Limitation and future work are presented	Yes
Reference format is consistent.	Yes
Main text >= 8 pages (not include references and appendix)	Yes