DIFFMEDDINOV3: A DIFFUSION-DISTILLED MULTI-TASK MULTI-MODAL FOUNDATION MODELS FOR MEDICAL IMAGE SEGMENTATION

Hao Guo, Juan Jose Serrano Mora , Khondker Fariha Hossain

School of Informatics, Computing, and Cyber Systems, Northern Arizona University, Flagstaff, AZ

ABSTRACT

Medical image segmentation is one of the mainstream tasks for diagnosis and treatment planning. Recently, a series of vision foundation models and multimodal large language models have been widely used in medical image segmentation. In the state-of-the-art model named MedDinoV3[1], the authors adapted the newly released vision foundation model DinoV3[2] into medical segmentation by building a knowledge distillation architecture with gram anchoring and high-resolution adaptation. They achieved impressive performance. However, gaps and noises of knowledge distillation and incongruity of the modality still exist. To address these limitations, we construct a multi-task teacher-student architecture to capture the semantic gaps across modalities. To discard the noisy information and distill the valuable information from transferring the teacher model into the student model, we utilize a diffusion model to denoise. Our approach is evaluated in several medical image segmentation datasets, achieving a xxx improvement in xxx and a xxx improvement in xxx. Our proposed model shows improved segmentation accuracy to adapt the latest large foundation model to fine-grained clinic domains, outperforming existing state-of-the-art methods.

***Index Terms*—** Vision Foundation Models, Medical Image Segmentation, Knowledge Distillation, Diffusion Model

1. INTRODUCTION

Medical image segmentation plays a vital role in clinical practice, enabling the precise identification and separation of anatomical structures that support diagnosis and treatment planning. However, existing approaches still face challenges in effectively transferring knowledge across modalities and reducing noise in learned representations, particularly in unsupervised or multi-modal segmentation tasks. Recent vision foundation models, such as MedDINOv3 [1], have achieved impressive results by adapting the DINOv3 [2] architecture to medical image segmentation through high-resolution adaptation, self-distillation, and gram anchoring. Despite their success, these methods still suffer from representation noise during the distillation process and limited adaptability across different imaging modalities, restricting their effectiveness in complex segmentation scenarios. To overcome these challenges, diffusion-based and distillation-driven foundation models have been explored for medical imaging [3, 4, 5]. Diffusion models enable progressive denoising and refinement of feature representations, enhancing the quality and stability of learned knowledge. We propose a method called XXX, which integrates a multi-task learning architecture into a teacher–student knowledge distillation framework [6, 7]. This design allows the model to capture both shared and task-specific representations across medical imaging modalities, improving feature transfer and generalization. Our approach is evaluated in several publicly available datasets, including BraTS (2018–2021) [8], AMOS22 [9], FeTS [10], and ISLES [11], covering modalities such as MRI and CT and anatomical regions such as the brain, abdomen, and ischemic lesions. Each dataset provides voxel-wise segmentation masks with expert annotations, enabling comprehensive evaluation of model generalization across modalities and tasks. Our model demonstrated XXX and XXX improvements in XXX and XXX compared to baseline implementations. In particular, the segmentation of challenging structures showed consistent performance gains, highlighting the robustness and adaptability of the model between modalities.

2. RELATED WORKS

Over five years, medical image segmentation shifted from CNNs to foundation/promptable models, but surveys note gaps in reasoning, generalization, and label efficiency [7, 12].

2.1. CNN-Based Methods

CNNs like U-Net [10] and FCNs [13] are reliable. U-Net++ [14] bridges gaps with dense skips, limited locally. Attention U-Net [15] focuses key areas, compute-heavy. ResUNet [16] and R2U-Net [17] stabilize gradients, miss long-range. DoubleU-Net [18] stacks encoders for depth, complex. PraNet [19] excels in polyps, domain-bound. KiU-Net [20] handles small structures, param-heavy. CNNs weak globally; self-attention [21, 22] helps slightly, spurring Transformers.

Thanks to Northern Arizona University for funding.

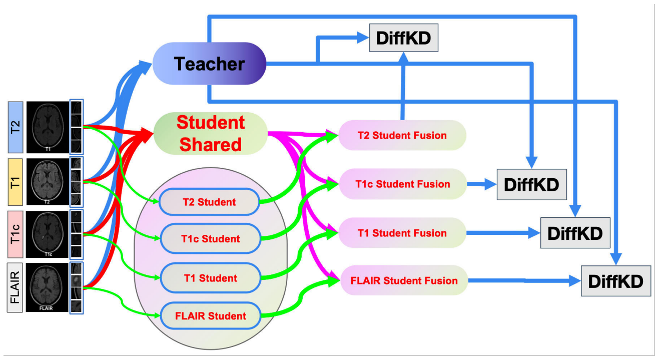

Fig. 1. Architecture of multi-task multi-modal knowledge distillation

2.2. Transformer-Based Approaches

Transformers [23] capture long-range ties. ViT [8] strong, lacks seg hierarchy. TransUNet [24] hybrids CNNs, bulky. Swin-Unet [25] efficient, window-sensitive. DS-TransUNet [26] fuses Swins, training-heavy. AA/DA-TransUNet [27, 28] sharpen focus, weak multi-modal.

SAM [29] zero-shot via prompts, poor medically. SegGPT [30] contextual, no domain fit. STU-Net [31] scales with pre-training, data-hungry. MedSAM [32] prompt-reliant; SAM Adapter [33] fast-tune, overhead; AutoSAM [34] no-prompt, weaker hard cases; SAMed [35] LoRA [36] efficient, overfit risk.

2.3. Foundation Models & KD in Multi-Modal

MedDINOv3 [1] adapts DINOv3 [2] via multi-scale tokens and CT pretraining, matches SOTA, but CT-limited and compute-heavy. One-Prompt [37] uses one clinician prompt for many unseen datasets, strong zero-shot but prompt-dependent. SEG-SAM [38] adds semantics, inconsistent. SAM.MD [39] good interactively, boundary weak. MAdapter [40] fuses image-text, tops tasks, risks noise. MCPL [41] aligns VLM, prompt-bound.

KD/diffusion cut labels: Task-Specific KD [6] distills VFM to small model via LoRA/synthetic data, beats low-data baselines. DiffuSeg [3] refines via diffusion, unstable. MoVE-KD [4] mixes encoders, reduces conflict. ClinKD [5] cross-modal with curriculum, SOTA multi-task.

Our diffusion-distilled multi-task cleans teacher noise, aligns modalities for robust cross-dataset.

3. METHODOLOGY

3.1. Model Architecture Overview

In this section, we will describe DiffMedDINOv3 model specifically with each part. The multi-task multi-modal knowledge distillation part is shown in Figure1. The medical image segmentation part is shown in Figure3. Also for the details of diffusion model with knowledge distillation, it is shown in Figure2. The framework utilizes the DINOv3 foundation model as the teacher model, along with a set of shared modality student model and modality-specific student models via using Swin-Transformer. The diffusion-based knowledge distillation mechanism is to align student representations with features learned by teacher model.

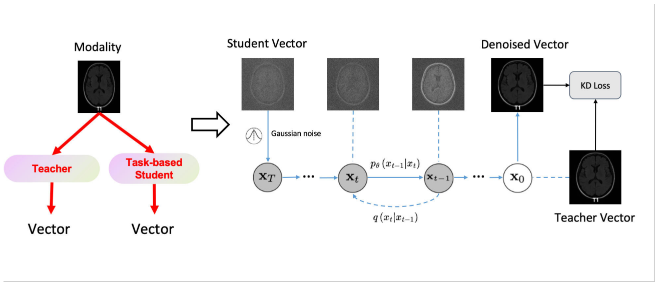

Fig. 2. The diffusion model in knowledge distillation

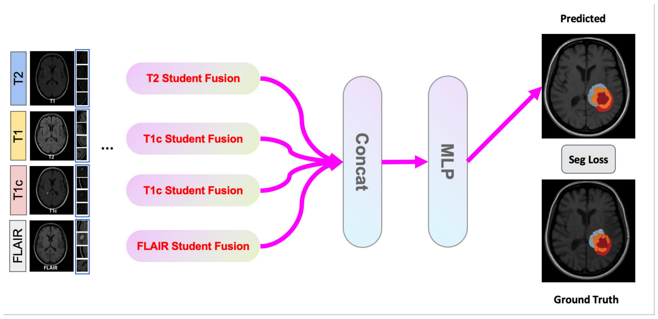

Fig. 3. Multi-task multi-modality model for medical image segmentation

3.2. Problem Formulation

Medical image segmentation with Multi-modal Magnetic Resonance Imaging (MRI) with different modalities is widely adopted in clinical practice. However, these modalities share similar morphology while keep their own modality-specific anatomical cues. In real clinical scenarios, modalities in MRI usually contains T1-weighted (T1), contrast-enhanced T1-weighted (T1c), T2-weighted (T2), and Fluid Attenuated Inversion Recovery (FLAIR) modalities. Formally, considering a multi-modal dataset $\{(\mathcal{X}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^{N}$. For the instance i, it consists of a collection of modalities $\mathcal{X}^{(i)} = \{\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \cdots, \mathbf{x}_M^{(i)}\}$ and its corresponding segmentation label $\mathbf{y}^{(i)}$. For the m-th modality in instance i, we use $\mathbf{x}_m^{(i)} \in \mathbb{R}^{H\times W\times D}$ to represent the corresponding image and the segmentation label $\mathbf{y}^{(i)} \in \{0, 1, 2, 3\}^{H\times W\times D}$ indicates the classification of each voxel which includes background, necrosis, edema and enhancing tumor.

As for the representation of teacher model $f_T(\cdot)$, it is extracted from a pretrained DINOv3[2] foundation model. As for the representation of student model, $f_S^s(\cdot)$ indicates modality-shared model, $f_S^m(\cdot)$ indicates the modality m's model. The fusion model $f_S(\cdot)$ encodes all above parts. Our goal is to learn a multitask multi-modal student model and use it into medical image segmentation.

3.3. Diffusion-based Knowledge distillation

3.3.1. Teacher model

We adopt a pretrained DINOv3 Vision Transformer[2] as the encoder of Teacher model. Given an input image $\mathbf{x}_m^{(i)}$ for the modality m and a patch size p, the model divides $\mathbf{x}_m^{(i)}$ into $N = \frac{H \times W}{p^2}$ patches, Each of them is transformed into a d-dimensional token representation via linear projection. The resulting patch-level matrix is $\mathbf{Z}_m^{(i)}(0) \in \mathbb{R}^{N \times d}$. Following the architecture of DINOv3, the backbone is a ViT[8] with L Transformer blocks. So the last layer token sequence is

$$\mathbf{Z}_m^{(i)}(L) = ViT\left(\mathbf{Z}_m^{(i)}(0)\right) \tag{1}$$

3.3.2. Student Model

The student model in **DiffMedDinoV3** consists of shared student model and modality-specific student model across the four MRI modalities. Since the medical image in different modality is high-resolution image comparing with natural image, so we use Swin-Transformer[9] to capture the high-resolution with the shift-window mechanism.

As for the shared student model $f_S^s(x)$, it is implemented with a multi-stage Swin-Transformer. Also for each modality, it has an independent Swin Transformer architecture. These architectures specialize in modeling modality-driven appearance. Mathematically, each modality has its own parameter set $f_S^m(x;\theta_m)$. For modality m and m', $\theta_m \neq \theta_{m'}$. Although each student model uses the same Swin Transformer structure, weights are not shared. Thus

$$\mathbf{A}_s^{(i)}(L) = SwinShared\left(\mathbf{Z}^{(i)}(0)\right) \tag{2}$$

$$\mathbf{H}_m^{(i)}(L) = SwinSpecific\left(\mathbf{Z}_m^{(i)}(0)\right) \tag{3}$$

Here, $m \in \{T1, T2, T1c, FLAIR\}$.

For the fused student model for each modality m, we firstly concatenate output from L'th shared student model $\mathbf{A}_s^{(i)}(L)$ and modality specific model $\mathbf{H}_m^{(i)}(L)$. Then we send the representation into a Swin-Transformer Block to get $\widetilde{\mathbf{H}}^{(i)}(L)$ with the same dimension as $\mathbf{A}_s^{(i)}(L)$ and $\mathbf{H}_m^{(i)}(L)$.

$$\widetilde{\mathbf{H}}_m^{(i)}(L) = SwinTBlock\left(Concat\left(\mathbf{A}_s^{(i)}(L), \mathbf{H}_m^{(i)}(L)\right)\right) \tag{4}$$

3.3.3. Knowledge Distillation

Large model is accomplished with the high requirement of large computation and memory. It would hard to do model training with the growth of volume of parameters. So knowledge distillation would boost the performance of efficient model (student model) via transfering the knowledge from larger model (teacher model). It means that we can use smaller model to get practically identical performance by larger model. However, due to the limited capacity or training recipe to learn truly valuable and decent features, student model usually is the noisy version. This would be detrimental for the student model shown in performance. Inspired by the success of previous generative tasks, we leverage diffusion models[11, 42] that can gradually remove the noise from an image or a feature.

As Figure 2 shows, we inject Gaussian noise in the forward diffusion process.

Let $t \in \{1, \cdots, T\}$ represents the diffusion timestep. Assume $\mathbf{u}_0^m = \widetilde{\mathbf{H}}_m^{(i)}(L)$ is the initial feature of student model under modality m.

In the forward process

$$q\left(\mathbf{u}_t^m|\mathbf{u}_0^m\right) = \mathcal{N}\left(\mathbf{u}_t^m|\sqrt{\overline{\alpha}_t}\mathbf{u}_0^m, (1-\overline{\alpha}_t)\,\mathrm{I}\right) \tag{5}$$

Thus the noisy representation via student model at time t is

$$\mathbf{u}_t^m = \sqrt{\overline{\alpha}_t}\mathbf{u}_0^m + \sqrt{1-\overline{\alpha}_t}\epsilon \tag{6}$$

In (6), $\epsilon \sim \mathcal{N}(0, \mathrm{I})$. In both (5) and (6), $\overline{\alpha}_t := \prod_{s=0}^{t} \alpha_s = \prod_{s=0}^{t}(1-\beta_s)$ is a notation for sampling $\mathbf{u}_t^m$ at arbitraty timestep with noise variance schedule β[11].

For the reversed process, we finally get the denoised representation of student model (also could be regarded as reconstructed teacher representation), the representation should be apprximate to representation of teacher model. The denoising model for modality m is D_θ^m. It is a lightweight feed-forward Transformer-like network. We express this with the formula as follows:

$$\hat{\mathbf{u}}_0^m = D_\theta^m\left(\mathbf{u}_t^m, t\right) \tag{7}$$

For the distillation target, we compute the error by comparing reconstructed teacher representation and real teacher representation and regard it as the reconstruction loss.

$$\mathcal{L}_{KD}^m = d\left(\hat{\mathbf{u}}_0^m, \mathbf{Z}_m^{(i)}(L)\right) \tag{8}$$

for the distance function d, we use simple MSE loss and KL divergence loss to compute the discrepancy of denoised student feature and teacher feature.

For M modalities, the diffusion KD loss is

$$\mathcal{L}_{KD} = \sum_{i=1}^{M} \mathcal{L}_{KD}^m = \sum_{i=1}^{M} d\left(\hat{\mathbf{u}}_0^m, \mathbf{Z}_m^{(i)}(L)\right) \tag{9}$$

3.4. Medical Image Segmentation

After the diffusion-based knowledge distillation, we can use a denoised fusion representation for each modality m. We then concatenate the four modality features as:

$$\hat{\mathbf{u}}_0 = Concat\left(\hat{\mathbf{u}}_0^0, \hat{\mathbf{u}}_0^1, \hat{\mathbf{u}}_0^2, \hat{\mathbf{u}}_0^3\right) \in \mathbb{R}^{N \times MC} \tag{10}$$

Here $\hat{\mathbf{u}}_0^0$ is the representation of $T1$, $\hat{\mathbf{u}}_0^1$ is the representation of $T2$, $\hat{\mathbf{u}}_0^2$ is the representation of $T1c$, $\hat{\mathbf{u}}_0^3$ is the representation of $FLAIR$.

Then we feed them to an MLP (multi-layer perceptron)-based segmentation decoder to predict the segmentation:

$$\hat{\mathbf{y}^{(i)}} = MLP\left(\hat{\mathbf{u}}_0\right) \in \mathbb{R}^{N \times n_{class}} \tag{11}$$

where n_{class} denotes the number of semantic class.

The segmentation loss is standard multi-class Dice and CE

$$\mathcal{L}_{seg} = \lambda_C \mathcal{L}_C\left(\hat{\mathbf{y}^{(i)}}, \mathbf{y}^{(i)}\right) + \lambda_D \mathcal{L}_D\left(\hat{\mathbf{y}^{(i)}}, \mathbf{y}^{(i)}\right) \tag{12}$$

3.5. Optimization

All tasks and modalities are jointly trained with the total loss.

$$\mathcal{L} = \mathcal{L}_{KD} + \mathcal{L}_{seg} \tag{13}$$

4. EXPERIMENTS

In this section, we describe experiment settings, datasets, evaluation metrics, quantitative comparison with baselines, ablation studies of DiffMedDINOv3.

4.1. Datasets

4.2. Experiment Setting

For each dataset, we follow the official training-testing split which has been finished by the organizers. All images are resized into 256×256 and normalized using the same mean and standard deviation parameters as in DINOv3[2].

We implement all experiments with Pytorch framework[43]. The models are optimized with AdamW[44] and learning rate of 1×10^{-4} along with a weight decay of 1×10^{-4}. We train the model for 100 epoches with a batch size of 4 on two NVIDIA A100 GPUs.

For the parameter settings in Teacher model, we follow the setting of DINOv3.

4.3. Evaluation Metrics

4.4. Quantitative Comparisson with baselines

4.5. Ablation studies

5. CONCLUSION

In this work, we introduced DiffMedDINOv3, a multi-task multi-modal knowledge distillation framework with diffusion model to denoise and segmentation task with a MLP-based decoder. By combining share student model and modality-specific student model and construct a diffusion model with representation from teacher model, the framework learn to denoise and recover teacher-aligned features, the student model overcomes architectural and contrast-related discrepancies that hinder classical feature-matching knowledge distillation. We conduct experiments on five medical image datasets. The results demonstrate the effectiveness of diffusion-based knowledge distillation in enhancing modality robustness. In the future, we plan to scale the framework to 3D diffusion transformers for volumetric distillation and apply diffusion-based distillation to other clinical tasks such as survival prediction.

6. REFERENCES

[1] Yuheng Li, Yizhou Wu, Yuxiang Lai, Mingzhe Hu, and Xiaofeng Yang, "Meddinov3: How to adapt vision foundation models for medical image segmentation?," *arXiv preprint arXiv:2509.02379*, 2025.

[2] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al., "Dinov3," *arXiv preprint arXiv:2508.10104*, 2025.

[3] Le Zhang, Fuping Wu, Katherine Bronik, and Bartlomiej W. Papiez, "Diffuseg," *IEEE Journal of Biomedical and Health Informatics*, 2025.

[4] Jiajun Cao, Yuan Zhang, Tao Huang, et al., "Move-kd," *arXiv preprint arXiv:2501.01709*, 2025.

[5] Hongyu Ge, Longkun Hao, Zihui Xu, et al., "Clinkd," *arXiv preprint arXiv:2502.05928*, 2025.

[6] Pengchen Liang, Haishan Huang, Bin Pu, et al., "Task-specific knowledge distillation," *arXiv preprint arXiv:2503.06976*, 2025.

[7] Fares Bougourzi and Abdenour Hadid, "Recent advances in medical imaging segmentation: A survey," *arXiv preprint arXiv:2505.09274*, 2025.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[9] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[12] Yibo Sun, Zhihao Chen, and Yuhang Chen, "Multimodal learning methods in medical imaging: A survey," *Digital Signal Processing*, vol. 152, pp. 105441, 2025.

[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.

[14] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis*, 2018, pp. 3–11.

[15] O. Oktay, J. Schlemper, L. Le Folgoc, et al., "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[16] Xiao Xiao, Shen Lian, Zhiming Luo, and Shaozi Li, "Weighted res-unet for high-quality retina vessel segmentation," in *ITME*, 2018, pp. 327–331.

[17] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "R2u-net: Recurrent residual convolutional neural network based on u-net for medical image segmentation," *arXiv preprint arXiv:1802.06955*, 2018.

[18] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "Doubleu-net," in *CBMS*, 2020, pp. 558–564.

[19] D.-P. Fan, G.-P. Ji, T. Zhou, et al., "Pranet: Parallel reverse attention network for polyp segmentation," in *MICCAI*, 2020, pp. 263–273.

[20] J. M. J. Valanarasu, V. A. Sindagi, I. Hacihaliloglu, and V. M. Patel, "Kiu-net: Accurate segmentation of biomedical images using over-complete representations," in *MICCAI*, 2020, pp. 363–373.

[21] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *CVPR*, 2018, pp. 7794–7803.

[22] J. Schlemper, O. Oktay, M. Schaap, et al., "Attention gated networks," *Medical Image Analysis*, vol. 53, pp. 197–207, 2019.

[23] Ashish Vaswani et al., "Attention is all you need," in *NeurIPS*, 2017.

[24] J. Chen, Y. Lu, Q. Yu, et al., "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[25] H. Cao, Y. Wang, J. Chen, et al., "Swin-unet," in *ECCV*, 2022, pp. 205–218.

[26] A. Lin, B. Chen, J. Xu, et al., "Ds-transunet," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, 2022.

[27] Y. Yang and S. Mehrkanoon, “Aa-transunet,” in *IJCNN*, 2022.

[28] G. Sun, Y. Pan, W. Kong, et al., “Da-transunet,” *Frontiers in Bioengineering and Biotechnology*, vol. 12, pp. 1398237, 2024.

[29] A. Kirillov, E. Mintun, N. Ravi, et al., “Segment anything,” in *ICCV*, 2023, pp. 4015–4026.

[30] X. Wang, X. Zhang, Y. Cao, et al., “Seggpt,” in *ICCV*, 2023, pp. 1130–1140.

[31] Z. Huang, H. Wang, Z. Deng, et al., “Stu-net,” *arXiv preprint arXiv:2304.06716*, 2023.

[32] J. Ma, Y. He, F. Li, et al., “Segment anything in medical images,” *Nature Communications*, vol. 15, no. 1, pp. 654, 2024.

[33] Yichi Zhang, Zhenrong Shen, and Rushi Jiao, “Segment anything model for medical image segmentation,” *Computers in Biology and Medicine*, p. 108238, 2024.

[34] Xinrong Hu, Xiaowei Xu, and Yiyu Shi, “How to efficiently adapt sam to medical images,” *arXiv preprint arXiv:2306.13731*, 2023.

[35] Kaidong Zhang and Dong Liu, “Customized segment anything model for medical image segmentation,” *arXiv preprint arXiv:2304.13785*, 2023.

[36] E. J. Hu, Y. Shen, P. Wallis, et al., “Lora: Low-rank adaptation of large language models,” *arXiv preprint arXiv:2106.09685*, 2021.

[37] Junde Wu, Jiayuan Zhu, Yueming Jin, and Min Xu, “One-prompt to segment all medical images,” *arXiv preprint arXiv:2305.10300*, 2023.

[38] Sheng He, Rina Bao, Jingpeng Li, Jeffrey Stout, and Li Yang, “Seg-sam,” *arXiv preprint arXiv:2412.12660*, 2024.

[39] Saikat Roy, Tassilo Wald, Gregor Koehler, et al., “Sam.md,” *arXiv preprint arXiv:2304.05396*, 2023.

[40] Xu Zhang, Bo Ni, Yang Yang, and Lefei Zhang, “Madapter,” in *MICCAI*, 2024.

[41] Pengyu Wang, Huaqi Zhang, and Yixuan Yuan, “Mcpl,” *IEEE Transactions on Medical Imaging*, vol. 43, no. 12, pp. 4224–4235, 2024.

[42] Jiaming Song, Chenlin Meng, and Stefano Ermon, “Denoising diffusion implicit models,” in *International Conference on Learning Representations*, 2021.

[43] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., “Pytorch: An imperative style, high-performance deep learning library,” *Advances in neural information processing systems*, vol. 32, 2019.

[44] Ilya Loshchilov and Frank Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.