PEMMA: Parameter-Efficient Multi-Modal Adaptation for Medical Image Segmentation

Nada Saadi, Numan Saeed, Mohammad Yaqub, and Karthik Nandakumar

Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE {
nada.saadi, numan.saeed, mohammad.yaqub, karthik.nandakumar}@mbzuai.ac.ae

Abstract. Imaging modalities such as Computed Tomography (CT) and Positron Emission Tomography (PET) are key in cancer detection, inspiring Deep Neural Networks (DNN) models that merge these scans for tumor segmentation. When both CT and PET scans are available, it is common to combine them as two channels of the input to the segmentation model. However, this method requires both scan types during training and inference, posing a challenge due to the limited availability of PET scans, thereby sometimes limiting the process to CT scans only. Hence, there is a need to develop a flexible DNN architecture that can be trained/updated using only CT scans but can effectively utilize PET scans when they become available. In this work, we propose a **p**arameter-**e**fficient **m**ulti-**m**odal **a**daptation (PEMMA) framework for lightweight upgrading of a transformer-based segmentation model trained only on CT scans to also incorporate PET scans. The benefits of the proposed approach are two-fold. Firstly, we leverage the inherent modularity of the transformer architecture and perform low-rank adaptation (LoRA) of the attention weights to achieve *parameter-efficient* adaptation. Secondly, since the PEMMA framework attempts to *minimize cross-modal entanglement*, it is possible to subsequently update the combined model using only one modality, without causing catastrophic forgetting of the other modality. Our proposed method achieves comparable results with the performance of early fusion techniques with just 8% of the trainable parameters, especially with a remarkable +28% improvement on the average dice score on PET scans when trained on a single modality.

Keywords: Multi-modal Adaptation · Low-rank Adaptation · Parameter-Efficiency · Cross-modal Entanglement · 3D Medical Image Segmentation

1 Introduction

In clinical practice, precisely identifying cancerous tumors through various imaging techniques presents a significant challenge due to the intricate nature of the disease. The necessity to interpret numerous imaging modalities complicates and lengthens the diagnostic and prognostic process, often making it arduous and monotonous. Consequently, there is a critical need for the advancement of Deep Neural Network (DNN) models that can automate this process. Moreover, there is a need to refine the integration of diverse modalities, mirroring the nuanced fusion approach instinctively employed by oncologists, thereby substantially enhancing accuracy and minimizing the intra- and inter-observer variability.

It is well-known that integration of both CT and PET imaging modalities using DNNs can significantly enhance the accuracy of tumor segmentation [8]. This can be attributed to the comprehensive evaluation of anatomical structures by the CT scans and the ability of PET to capture metabolic activity. However, the availability of PET scans may be restricted in practice due to cost and clinical necessity. Joint processing of these two imaging modalities has been extensively studied in the literature [11], [6]. These approaches involve either merging the modalities at the input level as channels [3] or at the output level after independent processing of each modality [9]. While the former approach assumes constant availability of both modalities, the latter approach doubles the number of trained parameters [12]. Hence, there is a need for more sophisticated method for jointly handling CT and PET scans.

In terms of DNN architecture used to process CT and PET modalities, several models such as Convolutional Neural Networks (CNNs), Graph Neural Networks, and transformers have been studied in the literature [10]. While CNN-based models have demonstrated impressive performance in various applications, opting for transformer-based models for segmentation tasks offers several advantages. Specifically, due to their unparalleled ability to capture long-range dependencies and intricate patterns within the data, transformer based models can lead to more refined and accurate outcomes. In this context, the UNETR model [4] marks a significant departure from traditional CNNs and uses the Vision Transformer (ViT) architecture. In general, the image encoder in a transformer-based segmentation model converts the given image $\mathbf{x}$ into a set of N patch tokens $\mathcal{T}_0 = \{\mathbf{t}_1, \cdots, \mathbf{t}_N\}$ via a patch embedding layer $\mathcal{E}_{\theta_{\mathrm{PE}}}$, where θ_{PE} denotes the patch embedding parameters. These tokens are processed through a sequence of L identical self-attention transformer blocks, where the operation of each block can be expressed as $\mathcal{T}_\ell = \mathcal{G}_{\theta_\ell}(\mathcal{T}_{\ell-1})$, where θ_ℓ represents the parameters of the ℓ^{th} transformer block, $\ell \in [1, L]$.

In this work, we address the problem of efficient multi-modal adaptation of a transformer-based tumor segmentation model. Suppose that a healthcare institution already has access to a pre-trained segmentation model based on CT scans. If new data containing both CT and PET scans becomes available, adapting the existing model to effectively utilize the PET information is not straightforward. We explore whether recent developments in parameter-efficient fine-tuning (PEFT) techniques such as Low-Rank Adaptation (LoRA) [5] can be leveraged to bridge the above gap. LoRA is one of the most PEFTs for adapting large language models like GPT-3 (with 175 billion parameters) to downstream tasks. It is based on the hypothesis that weight updates to the multi-head self-attention (MHSA) layer of a transformer during adaptation have a low intrinsic rank. Hence, LoRA injects trainable low-rank matrices in parallel to the attention layer while keeping the pre-trained model weights frozen. Based on LoRA, we propose a Parameter Efficient Multimodal Adaptation Approach

(PEMMA) that makes best use of the existing pre-trained model for CT and progressively integrates additional modalities (e.g., PET). Our contributions are:

- We propose a method that offers an efficient incorporation of new modalities into an existing model with minimal cross-modal entanglement.
- We demonstrate the feasibility of flexibly and efficiently fine-tuning the adapted model when only one modality is available.
- We show how our approach can effectively retain the knowledge learned in the previous steps when trained on new data.

2 Methodology

Problem Statement: Suppose that a pre-trained transformer-based tumor segmentation model $\mathcal{F}^{C}_{\Theta} : \mathcal{X}_C \rightarrow \mathcal{M}$ for CT scans is available, where $\mathcal{F}^{C}$ denotes the uni-modal (CT) model architecture with parameters Θ, $\mathcal{X}_C$ is the input space of CT scans, and $\mathcal{M}$ represents the output segmentation mask space. Our *primary goal* is to upgrade this model such that the adapted model $\widetilde{\mathcal{F}}^{CP}_{\Phi} : \mathcal{X}_C \times \mathcal{X}_P \rightarrow \mathcal{M}$ can utilize both CT and PET scans to achieve **better segmentation accuracy**, where $\widetilde{\mathcal{F}}^{CP}$ denotes the multi-modal (CT+PET) model architecture with parameters Φ and $\mathcal{X}_P$ is the input space of PET scans. We also have two secondary objectives: (i) the model adaptation must be **parameter-efficient**, i.e., the number of parameters in the multi-modal model (denoted as $|\Phi|$) must be close to that of the uni-modal model (denoted as $|\Theta|$), and (ii) the adaptation must **minimize cross-modal entanglement**, i.e., it should be possible to subsequently fine-tune the multi-modal model $\widetilde{\mathcal{F}}^{CP}_{\Phi}$ using only one modality (either CT or PET scans), without causing catastrophic forgetting of the other modality.

2.1 Standard Adaptation Methods

Early Fusion: The most common approach to train a multi-modal CT+PET segmentation model is to combine the two modalities as different channels of the same input, i.e., create a new multimodal input space $\mathcal{X}_{CP}$, and train a new model $\widehat{\mathcal{F}}^{CP}_{\Phi} : \mathcal{X}_{CP} \rightarrow \mathcal{M}$ using both the modalities [12]. When a transformer architecture (e.g., UNETR [4]) is used, this involves two main changes to the DNN architecture. Firstly, the uni-modal patch embedding (linear projection) layer that transforms the input data to a set of patch tokens is replaced with a new multi-modal patch embedding layer. Secondly, the direct skip connection between the input and the decoder needs to be modified to accommodate the increase in the number of channels. After making these two architectural changes, the entire model can be learned from the available CT+PET training data.

Specifically, let $\mathbf{x}_C \in \mathcal{X}_C$ be the CT image and $\mathcal{T}^{C}_{0} = \mathcal{E}_{\theta^{C}_{\text{PE}}}(\mathbf{x}_C)$ denote the set of CT patch tokens, where $\mathcal{E}_{\theta^{C}_{\text{PE}}}$ is the patch embedding layer of the uni-modal model $\mathcal{F}^{C}_{\Theta}$. Given the PET image $\mathbf{x}_P \in \mathcal{X}_P$, we first generate a combined image $\mathbf{x}_{CP} = [\mathbf{x}_C || \mathbf{x}_P] \in \mathcal{X}_{CP}$, where $||$ denotes channel-wise concatenation. Next, we train a new multi-modal patch embedding layer $\mathcal{E}_{\theta^{CP}_{\text{PE}}}$, which generates a new set

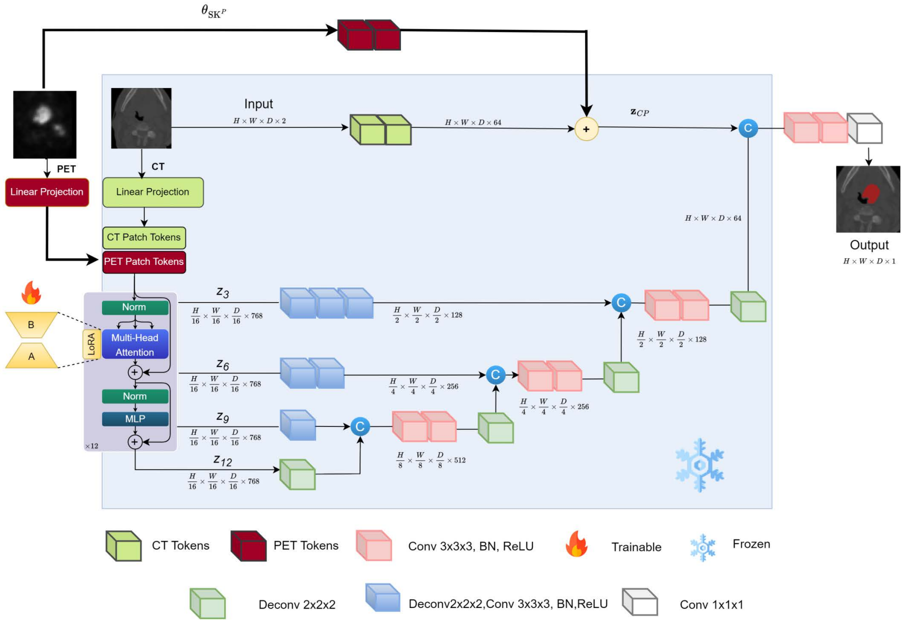
θ_{SK^P}
Input
$H \times W \times D \times 2$
$H \times W \times D \times 64$
$\mathbf{z}_{CP}$
PET
CT
Linear Projection
Linear Projection
CT Patch Tokens
PET Patch Tokens
$H \times W \times D \times 64$
Output
$H \times W \times D \times 1$
B
A
LoRA
Norm
Multi-Head Attention
Norm
MLP
×12
z_3
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 768$
$\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times 128$
$\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times 128$
z_6
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 768$
$\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 256$
$\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 256$
z_9
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 768$
$\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 512$
z_{12}
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 768$
C
CT Tokens
PET Tokens
Conv 3x3x3, BN, ReLU
Trainable
Frozen
Deconv 2x2x2
Deconv2x2x2,Conv 3x3x3, BN,ReLU
Conv 1x1x1

Fig. 1: ***Overview of our proposed architecture PEMMA:*** At the input level, we separate the path for CT and PET by adding the PET Skip Connection θ^P_{SK}. We freeze both the encoder and decoder part of the base UNetr model and introduce LoRA, after each ViT block (x12) as the only trainable layers.

of CT-PET patch tokens as $\mathcal{T}_0^{CP} = \mathcal{E}_{\theta^{CP}_{\mathrm{PE}}}(\mathbf{x}_{CP})$ that are passed to the subsequent transformer blocks in the encoder. Similarly, let $\mathbf{z}_C = \mathcal{S}_{\theta^C_{\mathrm{SK}}}(\mathbf{x}_C)$ be the output of the direct skip connection layer ($\mathcal{S}_{\theta^C_{\mathrm{SK}}}$) between the input and the decoder in the uni-modal model $\mathcal{F}^C_{\Theta}$. Again, we replace $\mathcal{S}_{\theta^C_{\mathrm{SK}}}$ with a new skip connection layer $\mathcal{S}_{\theta^{CP}_{\mathrm{SK}}}$, which outputs $\mathbf{z}_{CP} = \mathcal{S}_{\theta^{CP}_{\mathrm{SK}}}(\mathbf{x}_{CP})$.

The advantage of this adaptation approach is good parameter efficiency because only a minimal number of parameters need to be added to the patch embedding and input skip connection layers to account for the increase in the number of channels. However, the main limitation is that features from both CT and PET modalities get entangled in the combined model. Consequently, any attempt to fine-tune this model further using new data from only one modality (say CT only) will lead to the forgetting of the other modality. Since only minor architectural changes are involved in the adaptation, most of the parameters in the combined model can be initialized using the parameters of the pre-trained uni-modal model. However, for the newly introduced PET-related parameters in the combined model, there are three possible initialization strategies: (i) *random initialization* - initial weights are assigned stochastically, (ii) *zero initialization* - initial weights are set to zero, and (iii) *cross-modal initialization* - the weights of the other modality (CT) in the pre-trained model are re-used. The cross-

Table 1: HECKTOR Dataset description with the division of training and validation data across different centers.

Tasks	Centers	Train Samples	Val Samples	Pet/CT Scanner
Pre-training	CHUS CHUV CHUM CHUP	203	51	GeminiGXL 16 , Philips, Siemens Biograph mCT 40 TOF, Discovery D690 TOF
Multi-Modal Adaptation	MDA	152	44	Discovery HR, RX, ST, and STE
New Dataset 1	HGJ	39	15	Discovery ST, GE Healthcare
New Dataset 2	HMR	13	4	Discover ST, GE Healthcare

modal initialization approach can significantly accelerate the learning process and enhance model performance by guiding the PET channel with a pre-learned context of medical imaging, thereby facilitating a more effective CT-PET fusion.
Late Fusion: Another straightforward approach to incorporate a new modality such as PET is to train a completely new segmentation model $\mathcal{F}^P_\Psi : \mathcal{X}_P \to \mathcal{M}$ for PET scans, where $\mathcal{F}^P$ denotes the uni-modal (PET) model architecture with parameters Ψ, and combine the masks from the two models at the output stage. Let $M_C \in \mathcal{M}$ and $M_P \in \mathcal{M}$ be the masks produced by the uni-modal CT ($\mathcal{F}^C_\theta$) and PET ($\mathcal{F}^P_\Psi$) models, respectively. The combined mask can be computed as:

$$M_{CP} = w_C M_C + (1 - w_C) M_P, \tag{1}$$

where w_C is the weight assigned to the CT modality. While this method offers a great degree of flexibility in dealing with the dynamic availability of the two modalities during training and/or inference, it results in a two-fold increase in the number of parameters ($|\Phi| = |\Theta| + |\Psi|$) and may not provide optimal accuracy.

2.2 Proposed Adaptation Method: PEMMA

In light of the strengths and weaknesses of existing adaptation methods, we introduce a novel framework (see Fig. 1) for lightweight adaptation of a uni-modal model into a multi-model modal leveraging the inherent modularity of the transformer architecture. Our proposed framework, referred to as parameter-efficient multi-modal adaptation (PEMMA), is inspired by the concepts of visual prompt tuning (VPT) [7] and low-rank adaptation (LoRA) [5]. The PEMMA framework has three core components. Firstly, we introduce the new PET modality as a set of visual prompts (or context tokens) to the uni-modal model simply by adding a new patch embedding layer. Secondly, instead of fine-tuning all the parameters of the transformer encoder, we focus only on the attention layers and fine-tune these attention layers through LoRA matrices. Finally, instead of replacing the existing input skip layer in the uni-modal model, we add an additional uni-modal skip layer for PET that operates in parallel to the existing skip layer and the outputs of these two layers are linearly combined.

Given the PET image $\mathbf{x}_P$, we first add a new PET patch embedding layer $\mathcal{E}_{\theta^P_{\text{PE}}}$, which generates a new set of N PET patch tokens as $\mathcal{T}^P_0 = \mathcal{E}_{\theta^P_{\text{PE}}}(\mathbf{x}_P)$ that are passed to the subsequent transformer blocks in the encoder. Now, the operations of a transformer block can be represented as $\{\mathcal{T}^C_\ell, \mathcal{T}^P_\ell\} = \mathcal{G}_{\theta_\ell}(\{\mathcal{T}^C_{\ell-1}, \mathcal{T}^P_{\ell-1}\})$. Note that the above modification increases the number of tokens processed by the transformer encoder from N to $2N$. However, the decoder in the pre-trained uni-modal model can handle only N tokens. In order to avoid making changes to the decoder architecture, we allow only the N CT tokens from the intermediate transformer blocks ($\mathcal{T}^C_\ell$) to pass through to the decoder. Refer to Table 2 in the Appendix for Dimensionality Reduction results. It must be emphasized that the self-attention architecture ensures that the knowledge from the PET tokens gets distilled into the CT tokens, even though the PET tokens are ignored by the decoder layers. This is similar to VPT, where the learnable prompts only serve as the context and this contextual information gets distilled into the other tokens, even when the additional prompts are ignored in the end.

The MHSA parameters of a transformer block ℓ can be considered as a collection of four weight matrices denoted as $\{\mathbf{W}_{O,\ell}, \mathbf{W}_{Q,\ell}, \mathbf{W}_{K,\ell}, \mathbf{W}_{V,\ell}\}$. In LoRA, the updates to $\mathbf{W}_{Q,\ell}$ and $\mathbf{W}_{V,\ell}$ are decomposed into a pair of low rank matrices $\mathbf{A} \in \mathbb{R}^{r\times d}$ and $\mathbf{B} \in \mathbb{R}^{d\times r}$, where r represents the rank of the two matrices. Let $h_{*,\ell}$ and $\tilde{h}_{*,\ell}$ be the input and output, respectively, of an attention layer in the ℓ^{th} block. Then, LoRA operation can be summarized as:

$$\begin{aligned}\tilde{h}_{Q,\ell} &= \mathbf{W}_{Q,\ell} h_{Q,\ell} + \alpha \mathbf{B}_{Q,\ell}\mathbf{A}_{Q,\ell} h_{Q,\ell} \\ \tilde{h}_{V,\ell} &= \mathbf{W}_{V,\ell} h_{V,\ell} + \alpha \mathbf{B}_{V,\ell}\mathbf{A}_{V,\ell} h_{V,\ell}\end{aligned} \tag{2}$$

where α is a fixed scalar. Note that parameter-efficiency is achieved by allowing $\theta_{\text{LoRA}} = \{\mathbf{A}_{Q,\ell}, \mathbf{A}_{V,\ell}, \mathbf{B}_{Q,\ell}, \mathbf{B}_{V,\ell}\}_{\ell=1}^{L}$ as the only learnable parameters and freezing the rest of the parameters in the transformer encoder.

Finally, in contrast to the early fusion approach where $\mathcal{S}_{\theta^C_{\text{SK}}}$ is replaced with $\mathcal{S}_{\theta^{CP}_{\text{SK}}}$, we leave $\mathcal{S}_{\theta^C_{\text{SK}}}$ untouched and introduce an additional parallel path $\mathcal{S}_{\theta^P_{\text{SK}}}$ for the PET image. Let $\mathbf{z}_P = \mathcal{S}_{\theta^P_{\text{SK}}}(\mathbf{x}_P)$ be the output of the additional direct skip connection layer introduced for the PET modality. In this case, the combined output of the input skip layer is $\mathbf{z}_{CP} = \mathbf{z}_C + \beta\mathbf{z}_P$, where β is the weight assigned to the PET modality. The main motivation for the introduction of new patch embedding ($\mathcal{E}_{\theta^P_{\text{PE}}}$) and input skip layers ($\mathcal{S}_{\theta^P_{\text{SK}}}$) for the PET modality (rather than replacing them as in early fusion) is to minimize cross-modal entanglement. With the introduction of these additional layers, it is possible to subsequently fine-tune the multi-modal model using only one of the modalities, without affecting the model's ability to handle the other modality.

Flexible Training and Inference Strategy: The PEMMA framework introduces three new parameters θ^P_{PE}, θ_{LoRA}, and θ^P_{SK}. When adapting the unimodal model to the multimodal scenario, both CT and PET training data is required and all the new parameters $\{\theta^P_{\text{PE}}, \theta_{\text{LoRA}}, \theta^P_{\text{SK}}\}$ are learned, while the parameters of the pre-trained unimodal model Θ are completely frozen. Thus, the parameters of the multi-modal model are $\Phi = \{\Theta, \theta^P_{\text{PE}}, \theta_{\text{LoRA}}, \theta^P_{\text{SK}}\}$ and $|\Phi|$

is only marginally higher than $|\Theta|$ (hence, parameter-efficient). Subsequently, if new data is available to update the multi-modal model, only θ_{LoRA} needs to be updated and all other parameters can be frozen. This allows the flexibility of fine-tuning the multi-modal model using one or both modalities (if a modality is unavailable, the corresponding input can be set to zero). Similarly, the multi-modal allows flexible inference - it can effectively utilize both modalities when they are available, but can also be applied to only a single modality (albeit with some degradation in the segmentation accuracy).

Table 2: Results for the different Adaptation Approaches performances on the Training and Inference Modalities. We compare the number of trainable parameters with the UNETR model where Φ= 92.58M params. We compute the Tumor, Lymph, and Average Dice scores for each experiment. C=CT;P=PET;CP=CT+PET.

Dataset	⟶	**Multimodal Adaptation**			**New Dataset 1**						**New Dataset 2**					
Train Modalities	⟶	**CP**			**C**			**CP**			**C**			**CP**		
Infer Modalities	⟶	**CP**	**C**	**P**	**CP**	**C**	**P**	**CP**	**C**	**P**	**CP**	**C**	**P**	**CP**	**C**	**P**
Late Fusion **(params=2 Φ)**	Tumor	0.69	0.37	0.43	0.67	0.45	0.32	0.40	0.36	0.23	0.68	0.43	0.05	0.47	0.46	0.24
	Lymph	0.64	0.51	0.51	0.61	0.56	0.21	0.48	0.46	0.27	0.56	0.52	0.32	0.30	0.34	0.11
	Avg	0.65	0.43	0.47	0.64	0.49	0.17	0.44	0.39	0.25	0.62	**0.48**	0.19	0.39	0.35	0.18
Early Fusion **(params=1.0043 Φ)**	Tumor	0.65	0.41	0.68	0.76	0.71	0.17	0.81	0.42	0.30	0.82	0.34	0.01	0.63	0.35	0.42
	Lymph	0.63	0.48	0.64	0.63	0.64	0.01	0.56	0.48	0.37	0.58	0.38	0.04	0.56	0.35	0.25
	Avg	**0.67**	0.43	0.65	0.70	0.68	0.09	0.69	0.43	0.31	0.07	0.36	0.03	0.60	0.35	**0.34**
PEMMA (Ours) **(params=0.08 Φ)**	Tumor	0.67	0.41	0.64	0.82	0.70	0.29	0.81	0.43	0.31	0.86	0.38	0.34	0.64	0.35	0.39
	Lymph	0.60	0.50	0.63	0.61	0.72	0.30	0.57	0.48	0.40	0.68	0.50	0.26	0.57	0.48	0.26
	Avg	0.63	**0.44**	**0.65**	**0.74**	**0.69**	**0.28**	**0.72**	**0.45**	**0.33**	**0.75**	0.43	**0.31**	**0.63**	**0.41**	0.32

3 Experimental Set-up

3.1 Dataset Description

Pre-processing: The dataset [1] is publicly available on the MICCAI 2022 HEad and neCK TumOR (HECKTOR) challenge website. In total, the dataset compromises 522 samples, with a breakdown provided in Table 1 1 detailing distribution across various centers and specifying the scanner types utilized for scan acquisition. The training enhancements applied to both scans involve extracting four random crops sized 96×96×96. These augmentations aim to diversify and represent the training data more comprehensively For further details on the pre-processing steps refer to Table 1 in the Appendix.

Implementation Details: Our approach runs on PyTorch version 2.1.0 utilizing the MONAI Library [2]. We train all models for a maximum of 18k steps and select the best model based on the highest average dice score on validation set. We use a learning rate of 1e-4 and weight decay of 1e-5 for all experiments and train using AdamW optimizer with a batch size of 2. Our entire pipeline runs on a single Nvidia A6000 RTX 48GB GPU.

4 Results and Discussion

Our experimental approach leverages four centers of the HECKTOR Dataset (CHUM,CHUV,CHUP and CHUS) to construct our initial pre-trained unimodal (CT only), as shown in Table 1 1. We then fine-tune both modalities in the multimodal adaptation setting, leveraging data from MDA center.

We test our method on both joint (CT+PET) and separate modalities. We compare our findings with the standard adaptation approaches defined previously: Early and Late. We observe that PEMMA yields results on par with the early fusion approach, yet it is **12x** more efficient. To adapt to new datasets, we introduce the CT and PET scans from two new centers (HGJ and HMR) and design two training scenarios that reflect real-world medical environments: one model trained on CT only (when PET data is missing), and another with both modalities available. PEMMA surpasses both early and late fusion techniques across the various adaptation and inference strategies, particularly in enhancing the extra-modality (PET) tuning, evidenced by a remarkable increase of **+19%** and **+28%** in Average Dice scores for new datasets 1 and 2, respectively.

These outcomes highlight LoRA's effectiveness in progressively integrating additional modality information while significantly reducing the need for comprehensive model retraining, cutting down trainable parameters by **92%**. Finally, the proposed approach can also be useful in a continual learning setting, where new tasks with different data distributions may be introduced sequentially. Since LoRA is highly parameter-efficient, only very few parameters($|\theta_{\text{LoRA}}| \ll |\Phi|$) need to be updated for each task. Hence, it becomes much more feasible to store all task-specific parameters in memory and achieve good forward transfer as well as non-negative backward transfer (thereby circumventing catastrophic forgetting). We, therefore, successfully achieve the parameter efficiency and minimization of the cross-modality entanglement goals.

5 Conclusion

In this study, we introduce a Parameter-Efficient Multimodality Adaptation method to enhance the proposed model's adaptability and efficiency across various data sources and modalities without substantial retraining or a major increase in the number of the model's parameters. This technique allows our model to retain past knowledge and further learn from new data modalities, improving its ability to process evolving multimodal information. Looking ahead, we aim to test this method with other medical imaging types, like MRI, and expand our research to include diverse datasets. Further evaluation is needed to assess the model's reliability and flexibility in a wider range of medical imaging scenarios, potentially creating more effective diagnostic tools. Finally, an interesting research avenue that could be built on our multimodality adaptation approach is to investigate Parameter-Efficient Multi-disease Adaptation to allow DNNs to adapt efficiently and effectively when trained on data from different diseases.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Andrearczyk, V., Oreiller, V., Boughdad, S., Rest, C.C.L., Elhalawani, H., Jreige, M., Prior, J.O., Vallières, M., Visvikis, D., Hatt, M., Depeursinge, A.: Overview of the hecktor challenge at miccai 2021: Automatic head and neck tumor segmentation and outcome prediction in pet/ct images (2022)
2. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., Nath, V., He, Y., Xu, Z., Hatamizadeh, A., Myronenko, A., Zhu, W., Liu, Y., Zheng, M., Tang, Y., Yang, I., Zephyr, M., Hashemian, B., Alle, S., Darestani, M.Z., Budd, C., Modat, M., Vercauteren, T., Wang, G., Li, Y., Hu, Y., Fu, Y., Gorman, B., Johnson, H., Genereaux, B., Erdal, B.S., Gupta, V., Diaz-Pinto, A., Dourson, A., Maier-Hein, L., Jaeger, P.F., Baumgartner, M., Kalpathy-Cramer, J., Flores, M., Kirby, J., Cooper, L.A.D., Roth, H.R., Xu, D., Bericat, D., Floca, R., Zhou, S.K., Shuaib, H., Farahani, K., Maier-Hein, K.H., Aylward, S., Dogra, P., Ourselin, S., Feng, A.: Monai: An open-source framework for deep learning in healthcare (2022)
3. Farag, S., IJzerman, N.S., Houdijk, M.P., Reyners, A., Arens, A., Grünhagen, D.J., Desar, I.M., Gelderblom, H., Steeghs, N., Geus-Oei, L.d.: Early response evaluation using 18f-fdg-pet/ct does not influence management of patients with metastatic gastrointestinal stromal tumors (gist) treated with palliative intent. Nuklearmedizin - NuclearMedicine **60**, 411–416 (2021). https://doi.org/10.1055/a-1542-6211
4. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H., Xu, D.: Unetr: Transformers for 3d medical image segmentation (2021)
5. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021)
6. Huang, H., Qiu, L., Yang, S., Li, L., Nan, J., Li, Y., Han, C., Zhu, F., Zhao, C., Zhou, W.: Vision transformer-based multimodal feature fusion network for lymphoma segmentation on pet/ct images (2024)
7. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning (2022)
8. Jintao Ren, Jesper Grau Eriksen, J.N., Korreman, S.S.: Comparing different ct, pet and mri multi-modality image combinations for deep learning-based head and neck tumor segmentation. Acta Oncologica **60**(11), 1399–1406 (2021). https://doi.org/10.1080/0284186X.2021.1949034, https://doi.org/10.1080/0284186X.2021.1949034, pMID: 34264157
9. Kadoury, S., Wood, B.J., Venkatesan, A.M., Dalal, S., Xu, S., Kruecker, J.: Accuracy assessment of an automatic image-based pet/ct registration for ultrasound-guided biopsies and ablations. SPIE Proceedings (2011). https://doi.org/10.1117/12.878067
10. Peng, J., Peng, L., Zhou, Z., Han, X., Xu, H., Lu, L., Lv, W.: Multi-level fusion graph neural network: Application to pet and ct imaging for risk stratification of head and neck cancer. Biomedical Signal Processing

and Control **92**, 106137 (2024). `https://doi.org/https://doi.org/10.1016/j.bspc.2024.106137`, `https://www.sciencedirect.com/science/article/pii/S1746809424001952`

11. Saeed, N., Sobirov, I., Al Majzoub, R., Yaqub, M.: Tmss: An end-to-end transformer-based multimodal network for segmentation and survival prediction. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. pp. 319–329. Springer Nature Switzerland, Cham (2022)
12. Wang, Y.e.a.: Deep learning based time-to-event analysis with pet, ct and joint pet/ct for head and neck cancer prognosis. Computer Methods and Programs in Biomedicine **222**, 106948 (2022). `https://doi.org/10.1016/j.cmpb.2022.106948`