Parameter-efficient multi-task adaptors for echocardiographic image analysis

Adrian Thür
Centre for Artificial Intelligence (CAI)
Zurich University of Applied Sciences (ZHAW)
Winterthur, Switzerland
thueradr@students.zhaw.ch

***Abstract*—Foundation models are currently leading to a paradigm shift in artificial intelligence (AI) from models that have been trained on broad data and can be adapted to many downstream tasks. This work applies the paradigm of pretraining with a pre-text task and building problem-specific adapters for various downstream tasks based on echocardiography input images with Low-Rank Adaptation (LoRA). The interpretation of echocardiography images remains challenging and relies on expert knowledge, highlighting an opportunity for AI to extract quantitative information for clinical decision-making. LoRA enables fine-tuning of task-specific parameters while retaining the full capacity of the pretrained model. For this purpose, Segformer, a transformer-based architecture, is pretrained on the EchoNet-Dynamic dataset and serves as the backbone for our adapter. Segformer shows an accurate segmentation of the left ventricle with a dice of 0.926. The adapter with LoRA outperforms a fully trained convolutional neural network (CNN) in cardiac ultrasound view classification with an accuracy of 0.988 and ventricular volume regression with an MAE of 19.622 in the CAMUS dataset. In left ventricle segmentation, the adapter exceeds the performance of a fully trained Segformer MiT-B0 and MiT-B2 architecture with a dice of 0.897. For age determination, the associated adapter could not outperform a fully trained CNN with an MAE of 14.627.**

***Index Terms*—Artificial intelligence, Echocardiography, Semantic segmentation, Deep learning, Transfer learning**

I. Introduction

Cardiovascular disease (CVD) is the most common cause of death worldwide. It covers a wide array of disorders, including diseases of the cardiac muscle and the vascular system such as coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. Identifying patients with a high risk for a CVD can prevent one from death. [1]–[3]

Failures in the left ventricle in particular would likely result in impairment of all other organ systems. The left ventricle is a part of the cardiovascular system that pumps oxygenated blood through the aortic valve to the entire body by contraction. Almost one in five people are affected by left ventricular hypertrophy (LVH), an overworking heart due to arterial hypertension. Left untreated, it impairs left ventricular diastolic function and increases the risk of serious heart disease or even death. [4]

Echocardiography, also known as cardiac ultrasound, is a non-invasive and harmless assessment modality to assess the functionality of the left ventricle in real-time. Echocardiography produces images of the heart called echocardiogram using sound waves. An image gets reconstructed based on the echoes from the high-frequency sound waves emitted by the transducer and provides information about the heart structures at work. [5]

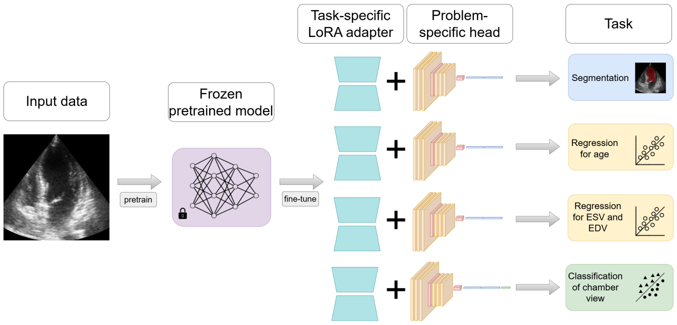

Figure 1. Each adapter consists of the same frozen pretrained model with task-specific LoRA weights and a problem-specific neural network.

Left ventricular ejection fraction (LVEF), the ratio of change in the left ventricular end-systolic (ESV) and end-diastolic volumes (EDV), is the central measure of left ventricular systolic function and to evaluate the cardiac contractility. Echocardiography information like LVEF is not only helpful for diagnosing the severity of CVD but also for determining the treatment strategy, prognosis, and treatment response evaluation. The calculation of parameters requires experienced practitioners to recognize even less common or subtly manifesting disorders, a challenge that automated image processing seeks to address by extracting quantitative measures. [6], [7]

The biplane method of disks (modified Simpson's rule) two-dimensional echocardiography method allows the assessment of LVEF. Compared to other methods, it relies less on geometric assumptions, since its calculation combines the apical four-chamber view and two-chamber view. To measure the LV volume the endocardial border is traced in both views. Then, the LV cavity is divided into a predetermined number of disks (usually 20) of the same height. The LV volume is obtained by summing up the disk volumes. This allows to further compute important measurement variables such as stroke volume $SV = EDV - ESV$ and ejection fraction $LVEF = \frac{EDV-ESV}{EDV} \star 100$. [6]

Convolutional neural network (CNN), a deep learning architectures, achieve accurate segmentation on medical images [8], [9]. However, recent advances with transformers (Vaswani et al. 2017 [10]) have shown promising results in computer vision tasks.

At present, the field of AI is undergoing a paradigm shift from models trained on broad data, called foundation models, which can be adapted to a wide range of downstream tasks. Their scale results in new emergent capabilities due to their implicitly induced behavior. Segmentation foundation models like Segment Anything (SAM) by Kirillov et al. (2023) [11] offer general-purpose segmentation capabilities, but the versatility of these models remains challenging due to the big difference between natural images and medical images [12]–[15].

Low-Rank Adaptation (LoRA) by Hu et al. (2021) [16] is a fine-tuning technique that has brought huge success in task-specific language models (LMs). This allows the model to acquire new skills or improve existing ones. LoRA is based on the hypothesis that the change in weights during fine-tuning has a low intrinsic rank. A low rank, the maximal number of linearly independent rows or columns of a matrix, means that it can be approximated by a small number of linearly independent columns. Linearly independent columns or rows are vectors that cannot be written as a linear combination of the others. Even a high-dimensional matrix can have a small rank caused by redundancy and therefore be represented by a linear combination of other columns. [16]

LoRA freezes the weights of the pretrained (foundation) model and injects trainable rank decomposition matrices into each layer of the transformer architecture. As a result, the model with LoRA retains the full capacity of the pretrained model and adds task-specific parameters across all transformer blocks using a fraction of the parameters of the pretrained model, resulting in less trainable parameters for the specific downstream task. A high-dimensional matrix can be represented as a product of two lower-dimensional matrices with rank decomposition. This way it is not necessary to retrain the entire model like in conventional fine-tuning methods. The results have shown that LoRA performs better than fine-tuned models although the fewer trainable parameters. The advantages are not only efficiency and lower hardware requirements but also that a pretrained model can be shared for multiple tasks by having multiple LoRA modules. [16]

Since the publication of LoRA in 2021 there have been new adaptations to the original method to make it more memory efficient based on quantization. However, the basic concept has not changed. [17], [18]

Wu et al. (2023) [19] evaluated the effect of LoRA on 17 medical image segmentation tasks across various image modalities, finding that LoRA fine-tuned models surpassed the performance of the traditional SAM and other state-of-the-art (SOTA) methods. Zhao et al. (2024) [20] presented with LoRA Land a web application with 25 task-specific LoRA fine-tuned Mistral-7B LLMs. Aside from the performance increase, the fact that it runs on a single NVIDIA A100 GPU with 80GB memory proves the cost effectiveness of employing multiple specialized LLMs over a single general-purpose LLM.

In this work, the author uses LoRA to adapt a pretrained model to semantic segmentation, image classification, and regression based on echocardiography input images. Specifically, the adapters aim to segment the left ventricle, classify the cardiac ultrasound view, and predict the patient's age along with the end-systolic and end-diastolic volume.

The main contribution is i) a custom pretraining of a transformer-based segmentation model, ii) the implementation and training of problem-specific adapters that leverage the pretrained model's prior knowledge for various tasks with LoRA, iii) the evaluation of these adapters on a second, lower quality dataset to assess their applicability to different downstream tasks, and iv) an investigation of the impact of the LoRA rank and different decompositions of the pretrained model on adapter performance.

II. Methods

The adapter with LoRA consists of a pretrained model as a backbone and a task-specific head. The transformer-based architecture Segformer by Xie et al. (2021) [21] is used for the pretrained model. Segformer includes a hierarchically structured transformer-based encoder that generates high-resolution coarse features and low-resolution fine features, as well as a multilayer perceptron (MLP) decoder. Compared to the vision transformer (ViT) for image classification proposed by Dosovitskiy et al. [22], Segformer does not need positional encoding to describe patch location information. This makes it possible that test resolution and training resolution do not have to match without interpolating the positional encoding. In Segformer a combination of depth-wise convolution and MLP called Mix-FFN Layer provides the positional information for the transformer. [21]

Similarly to ViT the input image of size $H \times W \times N_{channels}$, where $N_{channels}$ equals 3 for RGB or 1 for grayscale images, is converted into patches and fed to the encoder. However, in a Segformer architecture, overlapping patches preserve the local continuity around these patches. Hence, a kernel size of 7 (equal to the patch size), a stride of 4 and a padding of 3 are applied for stage 1 and a kernel size of 3, a stride of 2 and a padding of 1 are used for stage 2 to stage 4, resulting in 16'384 Patches for stage 1 or 65'536 patches for stage 2 to stage 4 with an input size of 512. Patch merging allows a hierarchical feature map F_i with a resolution of $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i$ where $i \in 1, 2, 3, 4$ are the 4 stages and C_i the embedding dimension. To improve the computational bottleneck, the authors introduced a reduction ratio R to reduce the length of the sequence of key K in the self-attention layer called efficient self-attention. Mix Transformer encoders (MiT, Mix-FFN and efficient self-attention) MiT-B0 to MiT-B5 have the same architecture but different sizes of complexity. [21]

The All-MLP decoder aggregates the information from the different encoder stages and so combines both local and global attention to create a comprehensive representation. The key to enabling such a simple decoder is that our hierarchical transformer encoder has a larger effective receptive field (ERF) than traditional CNN encoders. First, the decoder unifies the multiscale feature outputs of the encoder to the same channel dimension. The features are then upsampled to $\frac{H}{4} \times \frac{W}{4} \times C$, where C is the unified channel dimension. The features are then concatenated and fused to 1/4th of the channel dimension. Finally, the segmentation mask of size $\frac{H}{4} \times \frac{W}{4} \times N_{classes}$, where $N_{classes}$ is the number of classes, is predicted, and the prediction is interpolated to the input size. To reduce the complexity of the model, convolution layers are also implemented, which can be activated as hyperparameters to replace the original MLPs. [21]

The Segformer was pretrained on segmentation of the left ventricle using the EchoNet-Dynamic dataset, which contains 10'030 two-dimensional apical four-chamber echocardiogram videos in grayscale, each from a unique individual. The videos are randomly split into 7'465 for training, 1'277 for validation and 1'288 for testing. In each video, cardiologists labeled one frame for end-systole and one frame for end-diastole.

The CAMUS dataset (Cardiac Acquisitions for Multi-structure Ultrasound Segmentation) is used to evaluate the performance of LoRA adapters. It contains 2D echocardiographic images with two and four-chamber views of 500 patients including annotations for the left ventricle endocardium, the myocardium and the left atrium. The quality of the images ranges from wide variability to the fact that there was no preselection and should reflect clinical realism. The challenges are the different settings of acquisition, occlusions, and changes in the cardiac ultrasound view. The voxel spacing, which indicates the physical space for one pixel, is equal in all input images and can be ignored in tasks like regression. [8]

To create adapters for different tasks with LoRA we used LoraConfig from Parameter-Efficient Fine-Tuning (PEFT) by Huggingface. This library allows us to efficiently adapt our pretrained custom model to various downstream applications. In addition, it enables us to tune multiple hyperparameters. The most important ones among all are rank to set the rank of the decomposition matrix, target_modules to define the modules where LoRA should be injected and modules_to_save to set the trainable layers without LoRA.

To reduce the risk of human clerical mistakes, we create a LoRA adapter for the regression of the left ventricular EDV and ESV. When preprocessing the CAMUS dataset, the calculated LVEF is compared with the one from the dataset for a self-check whether ESV and EDV have been calculated correctly.

The adapters aim to make use of the comprehensive pretraining on the qualitatively higher and larger dataset. Therefore, one adapter is for the segmentation to perform a classic transfer learning on the CAMUS dataset. In this case, pretraining and fine-tuning are related problems. So, a third classification adapter should show the effect for another non-related task. This adapter attempts to classify whether it is an apical four-chamber view or a two-chamber view, which is supposed to help cardiologists acquire well-recognizable echocardiography images.

Controversial statements exist in echocardiographic studies for age-related cardiac values ranging from increase, in particular in women, to lack of statistical difference or linear correlation in LVEF with advancing age. Data obtained in a study show that RVEF and LVEF in children with normal hearts are similar to those in adults with normal cardiovascular systems, with an EF approximately in the range of 55% to 75%. The CAMUS dataset shows similar LVEF for all ages of patients. (see Figure 2) [23]–[26]

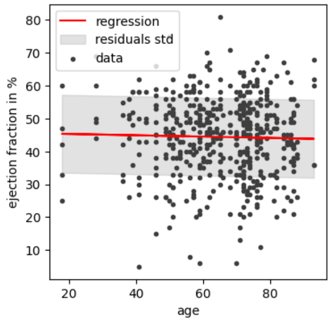

Figure 2. EF data with corresponding linear trend and residual standard error for each age of patients from the CAMUS dataset.

According to studies, ESV and EDV should decrease with advanced age, which implies increased EF. However, this effect is also not clearly reflected in the CAMUS dataset. The large standard deviation indicates that the observed data show a large variance around the mean volume for certain age groups of patients. Figure 3 shows many outliers with a higher volume than the median of 50 for ESV and 92.5 for EDV, which is also confirmed by a boxplot. [25]

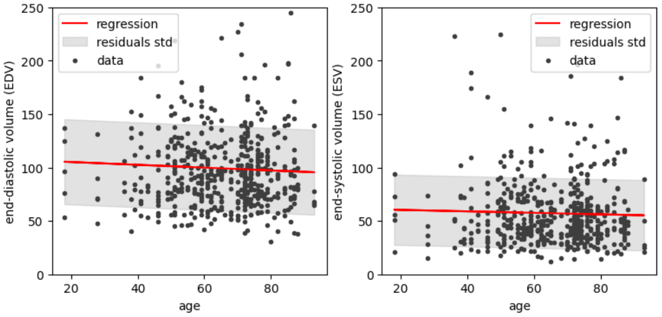

Figure 3. ESV and EDV data with corresponding linear trend and residual standard error for each age of patients from the CAMUS dataset.

Heart age and the resulting heart age gap are another metric to communicate a risk to a patient that is easy for the patient to understand. For the reasons mentioned, it is difficult to determine age directly using the EF, EDV, and ESV values. Therefore, in this paper we propose another adapter for age determination based on echocardiography input images. Regression of age is still difficult because the CAMUS dataset is biased towards older patients with a median of 67 years and only 94 patients with an EF between 55% and 75% are left.

III. Results

To compare the performance of the Segformer architecture, a UNet model, a successful architecture for a wide range of medical applications, is used as a baseline. The Segformer architecture outperforms the UNet on the EchoNet-Dynamic dataset, achieving a dice of 0.917 with a dice of 0.924 without having applied much hyperparameter tuning and augmentation techniques. After only 20 training epochs in 20 training hours, a dice of 0.926 could already be achieved with an elastic transformation. As there are no signs of overfitting, the result could be improved by longer training. Replacing the convolutional layers with linear layers in the embedding layer, the transformer layers or mask prediction layers do not gain any improvement. With MiT-B2, a more complex Segformer architecture, we achieve a similar dice of 0.925.

In this work, post-processing strategies such as connected component analysis or computer vision techniques like opening or closing are not used to further improve the performance, since the focus of this work is on the feasibility of task-specific adapters with LoRA. For the same reason, preprocessing is also kept to a minimum, such as input normalization. To overcome the challenge of different side ratios and sizes of input images, a center crop is applied to get the same side ratios and afterwards the images are resized to the same size.

In order to compare the results with the authors of the CAMUS dataset Leclerc et al. (2019) [8], all image qualities were used for training, were tested once with all image qualities and once without poor image quality. This corresponds to around 19% of images with poor quality and in absolute numbers to 10 out of 50 test images.

Apart from the hyperparameters i) batch size, ii) learning rate and iii) criterion, the author also does some experiments with the iv) composition of the adapter, v) different pretrained model complexities, and vi) different rank sizes. Playing with the composition of the adapter is an attempt to find out whether the entire pretrained model has an effect as a backbone or whether the encoder alone is sufficient.

It should be noted that the training time takes 30 minutes to load the data, which was not deducted from the training time. Training the problem-specific adapter head alone is the fastest method in all cases, which was to be expected, as all adapters consist of LoRA weights and the problem-specific adapter head. In addition, a LoRA rank of 1 is too low, and a rank of 8 is too high in all scenarios. Some experiments work best with a learning rate scheduler where cosine annealing was used. In 3 of the 4 problems, the adapter with LoRA weights could outperform the fully trained adapter head.

A. *Classification for the cardiac ultrasound view*

The problem-specific adapter head is a CNN with 2d convolution followed by batch norm, ReLU activation, dropout layer, and MaxPooling for 32, 64 and 128 hidden dimensions. In the fully connected network, ReLU is also used as the activation function, followed by a dropout layer for 1024, 512, and 256 hidden dimensions. Cross-entropy loss was used for the criterion and Adam for the optimizer. All experiments are trained over 200 epochs.

The adapter with LoRA outperformed the fully trained adapter head with an accuracy of 0.988 on the test images without poor image quality. In addition, the author tested whether the fully connected network is complex enough to classify the cardiac ultrasound view. The fully connected network with the full Segformer architecture as backbone outperformed the fully trained adapter head, but not all CNN adapter heads. The segmentation weights of the Segformer provide further useful information for classification. No improvement could be achieved with the more complex MiT-B2 architecture. The results were better when the LoRA weights were injected into all layers instead of just to the self-attention layers. (see Table I)

B. *Regression for the ventricular volume*

The problem-specific head is also a CNN and consists of a 2d convolution followed by batch norm, ReLU activation, and MaxPooling for 32, 64 and 128 hidden dimensions. Dropout was not beneficial for this problem. Huber loss, which combines the advantages of L1 loss and mean squared error (MSE) loss, was used as a criterion. All experiments are trained over 100 epochs with an Adam optimizer.

For this problem, the adapters with the LoRA weights outperform the fully trained adapter head as well. In addition, performance could be further improved with the MiT-B2 encoder architecture as backbone and LoRA weights only to the self-attention layers. (see Table II)

Figure 4 shows that the adapter produces biased predictions with R^2 of 0.75, resulting in overestimates for smaller volumes, as is the case with ESV, and underestimates for larger volumes, in the case of EDV.

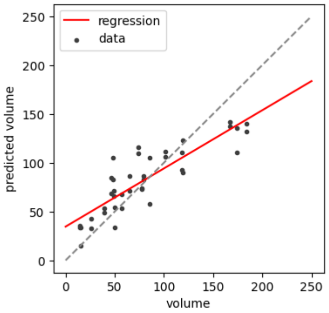

Figure 4. Evaluation of the test prediction of the LoRA adapter with the MiT-B2 backbone and LoRA weights injected into the self-attention layers.

C. *Regression for the age*

Regression for the age uses the same problem-specific adapter head, criterion, optimizer, batch size, and number of epochs as regression of ventricular volume. The only difference turned out to be the learning rate. The experiments worked best with a learning rate of 0.01.

For age determination, the LoRA adapters come close, but cannot outperform the fully trained adapter head. The MiT-B2 architecture does not have any improving effect in this case. For this problem, it worked better if the LoRA weights were injected only into the self-attention layer. (see Table III)

The test results of the best performing adapter reveals a bias with R^2 of 0.18, overestimating outcomes for younger ages and underestimating them for higher ages. The adapter predicts almost constant values that deviate from the mean of 65 years. (see Figure 5)

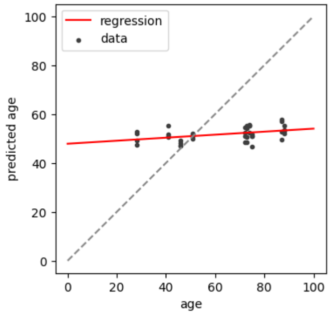

Figure 5. Evaluation of the test prediction of the best performing LoRA adapter with the full MiT-B0 pretrained model as backbone and LoRA weights injected into the self-attention layers.

D. *Left ventricle segmentation*

For the left ventricle segmentation, the Segformer decoder is used for the problem-specific adapter head with the only difference that concatenation and upsampling to a fourth of the input size is already done in the backbone. The experiments are trained over 100 epochs with cross-entropy loss and Adam optimizer.

The LoRA weights contribute to increased segmentation performance and slightly exceed the performance of the fully trained Segformer with MiT-B0 and MiT-B2 architecture. MiT-B2 for the pretrained model does not have an improving effect. For segmentation, it makes less of a difference whether the Lora weights are injected into all layers or only into the self-attention layers. Although only a fraction of the parameters of the pretrained model are trained, fine-tuning with Lora is slower than fully training a MiT-B0. (see Table IV)

The test results show that the majority of good results still have a few deviations at the edge of the chamber compared to the labels. The two of the three poorer results show that the labeled mask could not be interpreted correctly, resulting in grid-like lines in the mask. This could be improved using computer vision techniques. The worst segmentation result can be related to the abnormal shape of the left ventricle. (see Figure 6)

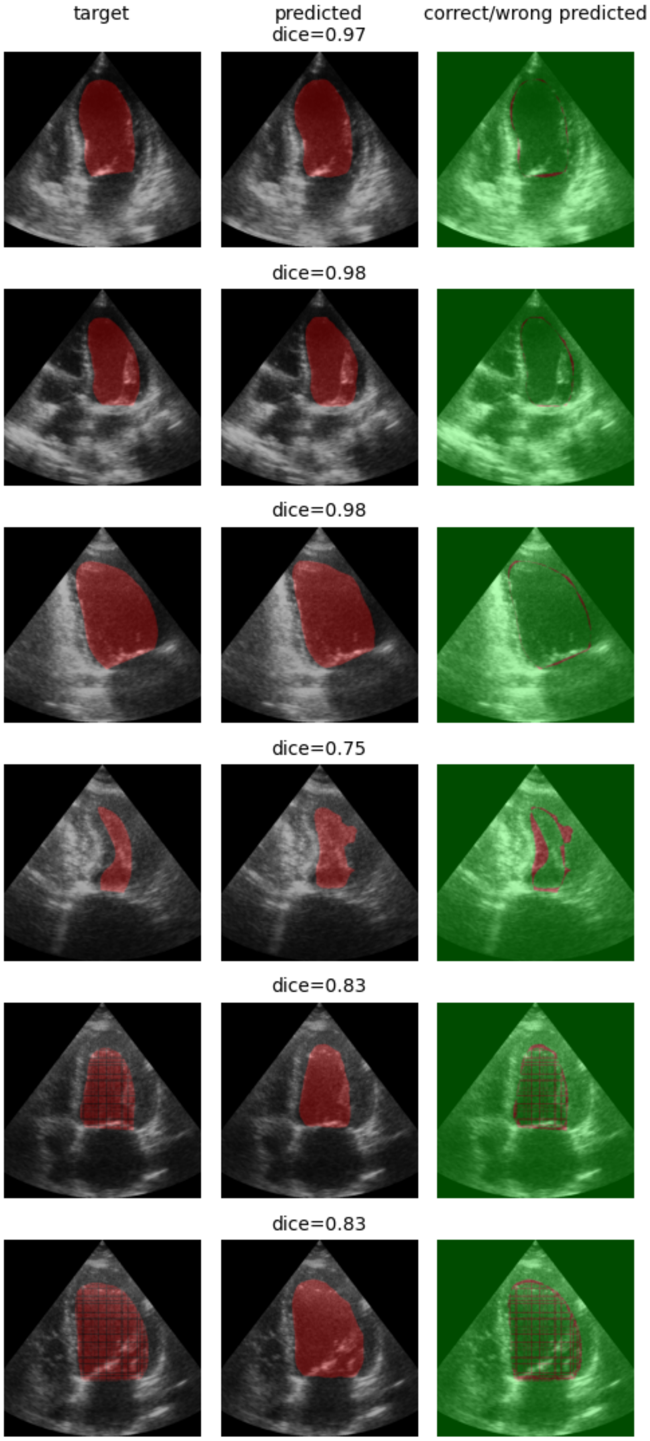

Figure 6. The 3 best and 3 worst test results from the segmentation adapter with LoRA.

IV. Discussion

This work utilized the paradigm of pretraining with a pretext task, followed by developing problem-specific adapters using LoRA to address semantic segmentation, regression and classification tasks based on echocardiography input images. The results have shown that although the adapter with LoRA could keep up with the fully trained Segformer the overall performance with a dice of 0.897 could be improved, as demonstrated by Leclerc et al. [8]. At the edge of the ventricle, there are still slight deviations between the labels and the test results. The results obtained must be analyzed with an expert and compared with the labels to determine further steps. In addition, an interrater agreement could put the results into

Table I

RESULTS FOR THE CLASSIFICATION ADAPTER

Experiment	layers with LoRA weights	rank	learning rate	test accuracy		trainable params	total params	training time [min]
				with poor images	***no poor images***			
backbone of MiT-B0	all	2	0.001	0.97	0.981	35’362’308	99’085’766	126
backbone of MiT-B0	all	4	0.001	0.965	0.962	36’124’774	99’848’232	123
backbone of MiT-B0	self-attention only	4	0.001	0.95	0.956	35’826’594	99’550’052	75
full MiT-B0[d]	all	4	0.001	**0.985**	**0.988**	136’412’550	300’681’058	137
full MiT-B0	self-attention only	4	0.001	0.955	0.956	539’112’164	1’106’121’192	109
backbone of MiT-B2[b]	all	4	0.001	0.955	0.962	39’758’950	265’848’840	409.5
backbone of MiT-B2[b]	self-attention only	4	0.001	0.93	0.956	38’659’490	264’749’380	191.5
full MiT-B0 with FCN only	all	4	0.001	0.97	0.981	1’531’084	568’446’704	129.5
full MiT-B0 with FCN only	self-attention only	4	0.001	0.93	0.944	1’490’178	568’405’798	80
fully train adapter head[c]	no LoRA	-	0.001	0.895	0.931	537’621’698	537’621’698	60.5

[a] “to” refers to a cosine anealing scheduler from the first-mentioned learning rate to the second-mentioned learning rate.

[b] Best performance with a batch size of 16 instead of 32

[c] Best performance with a batch size of 64 instead of 32

[d] Best performance with less complex adapter head with only 8, 16 and 32 hidden dimensions

Table II

RESULTS FOR THE VENTRICULAR VOLUME REGRESSION ADAPTER

Experiment	layers with LoRA weights	rank	learning rate	test mae		trainable params	total params	training time [min]
				with poor images	***no poor images***			
backbone of MiT-B0	all	4	0.001 to 0.0001[a]	32.88	30.255	71’122’789	169’844’262	99.5
backbone of MiT-B0	self-attention only	2	0.001	24.613	22.304	70’211’233	168’932’706	74.5
backbone of MiT-B0	self-attention only	4	0.001	23.98	22.27	70’824’609	169’546’082	71.5
full MiT-B0	all	4	0.001 to 0.0001[a]	29.963	26.988	1’077’467’565	2’182’791’088	155.5
full MiT-B0	self-attention only	4	0.001	25.701	23.401	1’077’426’659	2’182’750’182	127.5
backbone of MiT-B2	all	4	0.001	52.442	52.35	74’756’965	335’844’870	269
backbone of MiT-B2	self-attention only	4	0.001 to 0.0001[a]	**23.458**	**19.622**	73’657’505	334’745’410	184
fully train adapter head[b]	no LoRA	-	0.001 to 0.0001[a]	29.208	26.047	1’075’936’193	1’075’936’193	47

[a] “to” refers to a cosine anealing scheduler from the first-mentioned learning rate to the second-mentioned learning rate.

[b] Best performance with a batch size of 64 instead of 8

Table III

RESULTS FOR THE AGE REGRESSION ADAPTER

Experiment	layers with LoRA weights	rank	learning rate	test mae		trainable params	total params	training time [min]
				with poor images	***no poor images***			
backbone of MiT-B0	all	4	0.01	39.098	39.114	71’122’789	169’844’262	98.5
backbone of MiT-B0	self-attention only	2	0.01	15.332	14.627	70’211’233	168’932’706	73
backbone of MiT-B0	self-attention only	4	0.01	15.27	14.741	70’824’609	169’546’082	73.5
full MiT-B0[b]	all	4	0.01	45.45	45.532	539’152’813	1’106’161’584	153.5
full MiT-B0	self-attention only	4	0.001	15.385	14.717	1’077’426’659	2’182’750’182	127.5
backbone of MiT-B2	self-attention only	4	0.01	16.233	15.458	73’657’505	334’745’410	183
backbone of MiT-B2	all	4	0.01	23.649	23.56	74’756’965	335’844’870	266
fully train adapter head	no LoRA	-	0.001	**13.83**	**13.484**	1’075’936’193	1’075’936’193	64

[a] “to” refers to a cosine anealing scheduler from the first-mentioned learning rate to the second-mentioned learning rate.

[b] Best performance with dropout layers

Table IV
RESULTS FOR THE SEGMENTATION ADAPTER

Experiment	layers with LoRA weights	rank	learning rate	test dice		trainable params	total params	training time [min]
				with poor images	*no poor images*			
backbone of MiT-B0	all	2	0.0001 to 0.00001[a]	0.89	**0.897**	1’025’892	30’412’934	81.5
backbone of MiT-B0	all	4	0.0001 to 0.00001[a]	**0.891**	0.896	1’788’358	31’175’400	82.5
backbone of MiT-B0[b]	self-attention only	4	0.01 to 0.00001[a]	0.876	0.885	1’490’178	30’877’220	79.5
backbone of MiT-B2[c]	self-attention only	4	0.01 to 0.00001[a]	0.879	0.885	4’815’874	197’062’148	146.5
backbone of MiT-B2[c]	all	4	0.0001 to 0.00001[a]	**0.891**	0.897	5’915’334	198’161’608	233.5
fully train MiT-B0	no LoRA	-	0.001 to 0.0001[a]	0.883	0.89	29’387’042	29’387’042	70.5
fully train MiT-B2[c]	no LoRA	-	0.001 to 0.0001[a]	0.881	0.887	192’246’274	192’246’274	200.5

[a] “to” refers to a cosine anealing scheduler from the first-mentioned learning rate to the second-mentioned learning rate.

[b] Best performance with a batch size of 8 instead of 32

[c] Best performance with a batch size of 16 instead of 32

context. Classification seems to be a suitable problem for adapters with LoRA and achieved good performance with an accuracy of 0.988. With volume regression, an MAE of 19.622 was achieved, but the test results show a bias so that smaller volumes are overestimated and larger volumes are underestimated. A further step would be to investigate whether performance could be improved by training the regression of EDV and ESV as individual tasks. Currently, no gender or age information is included in the volume determination. This would have to be assessed for possible improvement. In age determination, the fully trained problem-specific network could not be outperformed but came close to the values. However, when examining the test results, it becomes apparent that constant values tend to be predicted and that lower ages are significantly overestimated and higher ages are significantly underestimated. For this regression problem, it would make more sense to have a dataset with more healthy patients, as we would like to estimate the heart age based on the assumption that the patient is healthy. Adapters with LoRA seem to be a suitable method for applications with small datasets. The performance achieved with the LoRA adapters on the echocardiography images would have to be examined in a further clinical analysis from a prospective study.

V. Conclusion

In this work, the author has used the paradigm shift in AI from models trained on broad data that can be adapted to various downstream tasks. The paradigm of pretraining with a pre-text task and building problem-specific adapters for different downstream tasks was applied with LoRA. It was shown that LoRA could be used to adapt a pretrained model for semantic segmentation, image classification, and regression based on echocardiography input image. In addition, Segformer has been shown to provide accurate segmentation results for echocardiography images.

Heart failure with preserved ejection fraction (HFpEF) is common among heart failure patients, for example from a disease such as arterial hypertension. Approximately 50% of patients with heart failure are classified as HFpEF. This is only one reason why it is important to examine as many factors as possible from the echocardiographic data. [27], [28]

Currently, only two frames for end-diastole and end-systole have been taken from each of the videos for pretraining. This means that a lot of information from the videos has not yet been included in the training. For this reason, the performance of the pretrained model could be further improved by a self-supervised technique such as DINO [29]. This way, the model can learn representations from unlabeled data with self-distillation.

The next step is not only to implement more adapters, but also to evaluate the entire video to detect abnormalities over time. The CAMUS dataset is not suitable for automated labeling of the end-diastolic and end-systolic time points, as only the smallest respectively biggest volume was included, and any presence of abnormalities was not considered.

Acknowledgment

This work was carried out as part of the requirements for the Master of Science in Engineering (MSE) in the in-depth thesis 1 at the Center for Artificial Intelligence (CAI) of the Zurich University of Applied Sciences (ZHAW). The author would like to thank Dr. Ahmed Abdulkadir for his expert support and constructive feedback throughout the research process and the CAI Institute for providing the server infrastructure.

References

[1] E. G. Nabel, “Cardiovascular disease,” *New England Journal of Medicine*, vol. 349, no. 1, pp. 60–72, 2003. DOI: 10.1056/NEJMra035098. [Online]. Available: https://www.nejm.org/doi/full/10.1056/NEJMra035098.

[2] W. H. Organization, *Cardiovascular diseases*, [Online]. Accessed: Dec. 2024. [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1.

[3] T. Gaziano, K. S. Reddy, F. Paccaud, S. Horton, and V. Chaturvedi, “Cardiovascular disease,” *Disease Control Priorities in Developing Countries. 2nd edition*, 2006.

[4] C. Clinic, *Left ventricular hypertrophy*, [Online]. Accessed: Dec. 2024. [Online]. Available: https://my.clevelandclinic.org/health/diseases/21883-left-ventricular-hypertrophy.

[5] M. K. Herbst, J. Velasquez, G. Adnan, and M. C. O'Rourke, "Cardiac ultrasound," *Cardiac Ultrasound*, pp. 1–151, Nov. 2022. DOI: 10.1201/9781315138800. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK470584/.

[6] A. Kosaraju, A. Goyal, Y. Grigorova, and A. N. Makaryus, "Left ventricular ejection fraction," *StatPearls*, Apr. 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK459131/%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6347358.

[7] S. H. Lee and J.-H. Park, "The role of echocardiography in evaluating cardiovascular diseases in patients with diabetes mellitus," en, *Diabetes Metab J*, vol. 47, no. 4, pp. 470–483, Jul. 2023.

[8] S. Leclerc, E. Smistad, J. Pedrosa, *et al.*, "Deep learning for segmentation using an open large-scale dataset in 2d echocardiography," *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2198–2210, 2019. DOI: 10.1109/TMI.2019.2900516.

[9] D. Ouyang, B. He, A. Ghorbani, *et al.*, "Video-based ai for beat-to-beat assessment of cardiac function," *Nature 2020 580:7802*, vol. 580, pp. 252–256, 7802 Mar. 2020, ISSN: 1476-4687. DOI: 10.1038/s41586-020-2145-8. [Online]. Available: https://www.nature.com/articles/s41586-020-2145-8.

[10] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention is all you need*, 2023. arXiv: 1706.03762 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1706.03762.

[11] A. Kirillov, E. Mintun, N. Ravi, *et al.*, *Segment anything*, 2023. arXiv: 2304.02643 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2304.02643.

[12] W. Ji, J. Li, Q. Bi, T. Liu, W. Li, and L. Cheng, "Segment anything is not always perfect: An investigation of sam on different real-world applications," *Machine Intelligence Research*, vol. 21, pp. 617–630, 4 Apr. 2023. DOI: 10.1007/s11633-023-1385-0. [Online]. Available: http://arxiv.org/abs/2304.05750%20http://dx.doi.org/10.1007/s11633-023-1385-0.

[13] 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. 102 918, 2023, ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2023.102918. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841523001780.

[14] Y. Huang, X. Yang, L. Liu, *et al.*, "Segment anything model for medical images?" *Medical Image Analysis*, vol. 92, p. 103 061, 2024, ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2023.103061. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841523003213.

[15] R. Bommasani, D. A. Hudson, E. Adeli, *et al.*, *On the opportunities and risks of foundation models*, 2022. arXiv: 2108.07258 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2108.07258.

[16] E. J. Hu, Y. Shen, P. Wallis, *et al.*, "Lora: Low-rank adaptation of large language models," *CoRR*, vol. abs/2106.09685, 2021. arXiv: 2106.09685. [Online]. Available: https://arxiv.org/abs/2106.09685.

[17] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36, Curran Associates, Inc., 2023, pp. 10 088–10 115. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf.

[18] Y. Xu, L. Xie, X. Gu, *et al.*, *Qa-lora: Quantization-aware low-rank adaptation of large language models*, 2023. arXiv: 2309.14717 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2309.14717.

[19] J. Wu, W. Ji, Y. Liu, *et al.*, *Medical sam adapter: Adapting segment anything model for medical image segmentation*, 2023. arXiv: 2304.12620 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2304.12620.

[20] J. Zhao, T. Wang, W. Abid, *et al.*, *Lora land: 310 fine-tuned llms that rival gpt-4, a technical report*, 2024. arXiv: 2405.00732 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2405.00732.

[21] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, *Segformer: Simple and efficient design for semantic segmentation with transformers*, 2021. arXiv: 2105.15203 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2105.15203.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, *An image is worth 16x16 words: Transformers for image recognition at scale*, 2021. arXiv: 2010.11929 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2010.11929.

[23] R. A. Hurwitz, S. Treves, and A. Kuruc, "Right ventricular and left ventricular ejection fraction in pediatric patients with normal hearts: First-pass radionuclide angiocardiography," en, *Am Heart J*, vol. 107, no. 4, pp. 726–732, Apr. 1984.

[24] Y. Singh, "Echocardiographic evaluation of hemodynamics in neonates and children," en, *Front Pediatr*, vol. 5, p. 201, Sep. 2017.

[25] M. Fiechter, T. A. Fuchs, C. Gebhard, *et al.*, "Age-related normal structural and functional ventricular values in cardiac function assessed by magnetic resonance," en, *BMC Med Imaging*, vol. 13, p. 6, Feb. 2013.

[26] C. Gebhard, B. E. Stähli, C. E. Gebhard, *et al.*, "Age- and gender-dependent left ventricular remodeling," en, *Echocardiography*, vol. 30, no. 10, pp. 1143–1150, Jun. 2013.

[27] R. S. Bhatia, J. V. Tu, D. S. Lee, *et al.*, "Outcome of heart failure with preserved ejection fraction in a population-based study," en, *N Engl J Med*, vol. 355, no. 3, pp. 260–269, Jul. 2006.

[28] M. S. G. Golla and P. Shams, "Heart failure with preserved ejection fraction (hfpef)," *Cardiovascular Manual for the Advanced Practice Provider: Mastering the Basics*, pp. 221–224, Mar. 2024. DOI: 10.1007/978-3-031-35819-7_21. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK599960/.

[29] M. Caron, H. Touvron, I. Misra, *et al.*, *Emerging properties in self-supervised vision transformers*, 2021. arXiv: 2104.14294 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2104.14294.

Code Availability

The code is available to the supervisor at https://github.com/adithuer/parameter-efficient-multi-task-adaptors-for-echocardiographic-image-analysis.