Automated Layer Selection for Efficient Fine-Tuning of Medical Image Segmentation Models

Denis Bolonin
240029096
Project Supervisor: Kit Mills Bransby
MSc Computer Science

***Abstract* - Fine-tuning large pre-trained models for medical image segmentation is a critical but computationally expensive task. Manual layer selection for parameter-efficient fine-tuning is often suboptimal and may lack scalability. This project aims to introduce an automated layer selection algorithm designed to optimize parameter efficiency during transfer learning for medical image segmentation. The proposed framework aims to overcome the limitations of manual approaches by adapting surgical fine-tuning and task-specific parameter freezing methods. A key contribution is the development of a multi-task preservation mechanism using adaptive parameter adaptation to minimize interference when fine-tuning for multiple related tasks, such as organ and tumor segmentation. The effectiveness of the algorithm will be validated against standard fine-tuning baselines on medical imaging datasets, using segmentation accuracy as key performance metric.**

***Keywords* - Automated layer selection, fine-tuning, medical image segmentation, computer vision**

I. INTRODUCTION

Deep learning models, particularly those leveraging advanced Transformer and convolutional architectures, have set new benchmarks in medical image segmentation. State-of-the-art models like CSWin-UNet [3] demonstrate remarkable accuracy by effectively capturing complex spatial hierarchies in medical scans. The standard paradigm for deploying these models involves transfer learning, where a model pre-trained on a large-scale dataset is fine-tuned for a specific downstream task. However, as these models grow in size and complexity, the process of full fine-tuning becomes too expensive in terms of computational resources and time, posing a significant barrier in data-limited and resource-constrained settings.

To address this challenge, various parameter-efficient fine-tuning methods have been developed. While these techniques reduce training costs by updating only a subset of parameters, early approaches often rely on simple heuristics like manual layer freezing or introduce architectural changes that may not be optimal. Recent works have shown more principled, automated approaches. For instance, Surgical Fine-Tuning [2] proposes identifying and preserving a core feature subspace to prevent catastrophic forgetting, while the Trainable Projected Gradient Method (TPGM) [5] learns layer-specific constraints to regularize updates automatically. These methods highlight a critical insight: the key to efficient adaptation lies in intelligently identifying what to update and how strong the updates should be.

Building on these principles, this work proposes an automated method to enhance the efficiency and robustness of fine-tuning for medical image segmentation. I will adapt and evaluate modern architectures, such as CSWin-UNet, within this framework. The core of my work is to develop an algorithm that automatically determines an optimal fine-tuning strategy, balancing high segmentation accuracy and computational efficiency.

A significant contribution of this project is the integration of a multi-task preservation mechanism, which is crucial for clinical workflows where a general model might be sequentially fine-tuned on related tasks (e.g., segmenting an organ and then a tumor within it). Inspired by the regularization techniques in methods like TPGM and Surgical Fine-Tuning, my method will incorporate adaptive parameter selection to isolate task-specific updates. This enables the model to retain knowledge from previously learned tasks, mitigating catastrophic forgetting and cross-task interference while adapting to new segmentation challenges.

The proposed method will be rigorously evaluated on public medical imaging datasets containing CT scans. Its performance will be compared against baseline approaches, including full fine-tuning, using established metrics such as the dice and Hausdorff distance for segmentation accuracy. This project seeks to provide a practical and robust solution for tuning large-scale vision models in resource-constrained environments, making advanced models more accessible and adaptable for clinical use.

II. LITERATURE REVIEW

This literature review examines three key areas of research that form the foundation for this project: efficient medical image segmentation architectures, fine-tuning methods that preserve important pre-trained knowledge, and automated approaches to layer-wise regularization. Together, these works provide the technical foundation and motivation for developing an automated fine-tuning framework specifically designed for medical image segmentation tasks.

A. CSWin-UNet for Efficient Medical Image Segmentation

Liu et al. (2025) developed CSWin-UNet to address a fundamental challenge in medical image segmentation: how to capture global context like Transformers do, but without their massive computational cost. Traditional CNN-based models like UNet work well but struggle with long-range dependencies, while Transformer models can capture global information but are computationally expensive. Previous attempts like Swin-UNet tried to bridge this gap with windowed attention, but still had limited interactions between different parts of the image.

CSWin-UNet's main innovation is the cross-shaped window attention mechanism. Instead of looking at square windows like previous methods, it processes information along horizontal and vertical strips simultaneously. This simple change dramatically expands how much context each part of the image can "see" without significantly increasing computational cost. The model also uses a content-aware upsampling method called CARAFE, which predicts how to reassemble features based on their content, leading to sharper and more accurate segmentation boundaries.

On the Synapse multi-organ CT dataset, CSWin-UNet achieved 81.12% Dice similarity coefficient, outperforming both CNN-based methods like UNet and other Transformer models like Swin-UNet. More importantly for this project, it accomplished this with only

23.57M parameters, making it one of the most efficient high-performing models available. However, even with these efficiency improvements, fine-tuning the entire model remains computationally expensive, especially for resource-constrained medical institutions.

B. Preserving Core Features via Surgical Fine-Tuning

The core problem that Lee et al. (2022) tackle is straightforward but critical: when you fine-tune a pre-trained model, you often destroy the valuable knowledge it learned during pre-training. Full fine-tuning gives you flexibility but can lead to overfitting, especially with small datasets. On the other extreme, just training a new classifier on frozen features (linear probing) is often too restrictive and doesn't adapt well to new tasks.

Their solution, called Surgical Fine-Tuning, is based on a key insight: not all learned features are equally important. They propose that a model's feature space can be divided into two parts - a "core" subspace containing essential, general-purpose features, and a "residual" subspace that can be safely modified. The trick is figuring out which is which.

Their method works by analyzing the pre-trained model's final classifier layer. Using singular value decomposition, they identify the most important directions that the model relied on for the original task. These directions define the core subspace that should be protected. During fine-tuning, they freeze the components of the feature extractor that align with this core subspace, while allowing updates to the orthogonal directions.

To make this practical, they developed Auto-Tune, which automatically learns how much to update each feature direction instead of requiring manual selection of which features to freeze. This automated approach consistently outperformed both full fine-tuning and linear probing across multiple benchmarks, with the biggest improvements coming in challenging scenarios with small datasets or large distribution shifts - exactly the conditions common in medical imaging.

C. Automated Layer-wise Regularization for Robust Fine-Tuning

While Surgical Fine-Tuning focuses on which feature directions to preserve, Tian et al. (2023) take a different approach with the Trainable Projected Gradient Method (TPGM). They discuss the same core problem - preventing fine-tuning from destroying useful pre-trained features - but do so by learning how far each layer should be allowed to deviate from its pre-trained initialization.

TPGM's key idea is to constrain each layer's weights to stay within a learnable "radius" of their original values. After each gradient update, if a layer's weights have moved too far from their initialization, they get projected back into an allowed region. The breakthrough is that these constraints aren't fixed - they're learned automatically through an optimization process.

In the inner optimization loop, the model weights are updated normally using training data. In the outer loop, the projection radii themselves are treated as learnable parameters and updated using a separate validation set. This means the model learns the optimal level of regularization for each layer, automatically balancing between preserving pre-trained knowledge and adapting to the new task.

The experimental results showed that TPGM significantly improves out-of-distribution robustness compared to standard fine-tuning, while being much more practical than methods requiring expensive hyperparameter searches. Unlike Surgical Fine-Tuning, which operates on feature directions, TPGM provides layer-wise control, offering a complementary approach to intelligent fine-tuning.

These three works provide complementary insights into efficient model adaptation. CSWin-UNet demonstrates that architectural efficiency is crucial for medical imaging applications, achieving strong performance with relatively few parameters. Surgical Fine-Tuning and TPGM show that intelligent parameter selection and regularization can preserve important pre-trained knowledge while enabling effective adaptation.

However, several gaps remain that this project aims to address. First, none of these works specifically tackle the sequential learning scenarios common in medical workflows, where a model might be fine-tuned on related tasks over time (e.g., organ segmentation followed by tumor segmentation within that organ). Second, the combination of efficient architectures like CSWin-UNet with principled fine-tuning methods hasn't been explored. Finally, medical imaging datasets have unique characteristics - small dataset sizes, high inter-patient variability, and critical accuracy requirements - that may benefit from specialized adaptation strategies.

This project will build on these foundations by developing an automated framework that combines the efficiency of modern architectures with the principled approach of methods like Surgical Fine-Tuning and TPGM, specifically designed for the challenges of medical image segmentation.

III. METHODOLOGY

The foundational model for this research is CSWin-UNet, a pure Transformer architecture that has demonstrated state-of-the-art performance in medical image segmentation. This model was selected because it effectively balances the global context modeling capabilities of Transformers with computational efficiency, making it an ideal testbed for developing automated fine-tuning methods for resource-constrained medical imaging applications.

CSWin-UNet addresses a fundamental limitation of standard Vision Transformers: their quadratic computational complexity makes them impractical for high-resolution medical images. The architecture employs CSWin Transformer blocks in both the encoder and decoder, featuring a cross-shaped window self-attention mechanism that computes attention along horizontal and vertical stripes simultaneously. This design significantly expands the receptive field for each token compared to traditional windowed attention approaches, enabling better capture of long-range dependencies without prohibitive computational costs.

The first component of my hybrid approach builds on Surgical Fine-Tuning, which addresses the fundamental trade-off in transfer learning: full fine-tuning risks destroying valuable pre-trained knowledge, while linear probing may be too restrictive for effective adaptation. Surgical Fine-Tuning decomposes the learned feature space into "core" and "residual" subspaces, where core features represent essential, generalizable knowledge that should be preserved.

The method identifies these critical feature directions by analyzing the pre-trained model's decision-making process. Using singular value decomposition on the final classification layer, it determines which feature directions were most important for the source task. During fine-tuning, the components of the feature extractor that align with this core subspace are frozen, while orthogonal directions in the residual subspace remain trainable. This

selective preservation allows the model to adapt to new tasks while maintaining its fundamental understanding of visual features.

The second component employs the Trainable Projected Gradient Method (TPGM), which takes a complementary approach to preserving pre-trained knowledge. Rather than identifying specific feature directions to freeze, TPGM learns how far each layer should be allowed to deviate from its pre-trained initialization through adaptive regularization.

TPGM constrains each layer's weights to remain within a learnable L2-norm radius of their original values. The key innovation is that these constraint radii are not fixed hyperparameters but are automatically learned through bi-level optimization. In the inner optimization loop, model weights are updated using standard gradient descent on the training loss. In the outer loop, the projection radii themselves are treated as learnable parameters and updated using a validation set to optimize the model's generalization performance.

After each weight update, if a layer has moved beyond its learned radius from initialization, the weights are projected back into the allowable region. This mechanism provides layer-wise control over the adaptation process, automatically determining which layers require more flexibility for the new task versus which should remain closer to their pre-trained state.

My proposed method combines these two approaches to leverage their complementary strengths. Surgical Fine-Tuning provides precise, theoretically grounded preservation of the most critical feature directions, while TPGM offers adaptive, data-driven regularization for the remaining parameters.

The integration operates as follows: First, the surgical analysis identifies and freezes the core feature subspace that represents essential knowledge from the pre-trained model. This creates a hard constraint preserving the most generalizable features. Second, for all trainable parameters - including both the residual subspace from the surgical modification and all other layers in the network - TPGM constraints are applied.

This dual-constraint approach allows the algorithm to maintain strict preservation of proven essential features while providing adaptive flexibility for other parameters. The TPGM component automatically learns appropriate regularization levels for each layer, balancing preservation of useful pre-trained knowledge with adaptation to the target task. The result is a fine-tuning method that combines explicit feature preservation with learned parameter-level constraints.

IV. EXPERIMENTAL SETUP

A. Datasets and Task Design

Three publicly available abdominal CT datasets were selected to create a realistic sequential learning scenario: Synapse (multi-organ segmentation), KiTS (kidney tumor segmentation), and LiTS (liver tumor segmentation). This progression mirrors real clinical workflows where models are first trained on general anatomy before specializing in specific pathologies.

The Synapse dataset serves as the foundation task, containing 30 CT scans with annotations for eight abdominal organs (aorta, gallbladder, left kidney, right kidney, liver, pancreas, spleen, stomach). KiTS focuses on kidney pathology with annotations for kidneys, tumors, and cysts, while LiTS targets liver pathology with liver and tumor annotations. To evaluate robustness to domain shifts, a modified Synapse-Blurred dataset was created by applying Gaussian blur to simulate common image quality degradation.

B. Segmentation Tasks

The evaluation follows a three-stage continual learning protocol:

- Task 1: Train on Synapse CT dataset (general abdominal anatomy)

- Task 2: Fine-tune Task 1 model on KiTS (kidney tumors)

- Task 3: Fine-tune Task 2 model on LiTS (liver tumors)

Each fine-tuning method (standard full fine-tuning, Surgical Fine-Tuning, TPGM, and the proposed hybrid) was evaluated through this sequence. To measure catastrophic forgetting, the final model was tested on all three datasets. Success requires not only good performance on the current task but also retention of previously learned capabilities.

Two critical baselines were established: (1) training from scratch on each dataset independently, and (2) standard full fine-tuning where all parameters are updated. Performance was measured using Dice Similarity Coefficient for overlap accuracy and 95th percentile Hausdorff Distance for boundary precision.

V. RESULTS AND ANALYSIS

A. Baseline Performance

Training from scratch established the upper bound for task-specific performance. On KiTS, this approach achieved DSC scores of 0.915 (kidney), 0.759 (tumor), and 0.703 (cysts) after 7 hours of training. On LiTS, it reached 0.895 (liver) and 0.407 (tumor) after 8 hours.

Standard fine-tuning demonstrated the classic efficiency-accuracy trade-off. For KiTS, fine-tuning the Synapse-pretrained model completed in just 55 minutes but produced inferior results: tumor DSC dropped to 0.625 and boundary accuracy worsened significantly (HD95 increased from 8.27 to 23.71). This suggests that general anatomical features did not effectively transfer to specialized lesion detection.

Conversely, on LiTS, fine-tuning proved beneficial. Training time reduced to 45 minutes while achieving better performance than training from scratch: liver DSC improved to 0.909 and tumor DSC increased to 0.477. This indicates that liver segmentation benefits more from general abdominal knowledge than kidney tumor detection.

B. Domain Shift Robustness

The domain shift experiment revealed important insights about knowledge preservation. The original Synapse model achieved an average DSC of 0.795 across all organs. After full fine-tuning on the blurred dataset, performance on the blurred test set dropped to 0.720, with particularly severe degradation for complex organs like the pancreas (0.75 to 0.64) and spleen (0.94 to 0.84).

Surgical Fine-Tuning proved more efficient and robust for this scenario. It completed adaptation in 30 minutes (half the time of full fine-tuning) while achieving comparable overall performance (DSC: 0.716 vs 0.720). More importantly, it dramatically better preserved boundary accuracy for specific organs - achieving HD95 scores of 1.93 (aorta) and 2.59 (spleen) compared to 59.48 and 34.08 for full fine-tuning. This demonstrates that preserving core feature directions prevents catastrophic degradation of edge detection capabilities.

C. Sequential Task Learning Results

The sequential learning experiments revealed the fundamental challenge of continual learning and highlighted the strengths and limitations of each approach.

Surgical Fine-Tuning Performance:

When adapting from Synapse to KiTS, Surgical Fine-Tuning achieved modest performance on the new task (tumor DSC: 0.610) but this was inferior to both training from scratch (0.759) and full fine-tuning (0.625). The rigid preservation of source-task features proved counterproductive for learning the specialized patterns required for tumor segmentation.

Knowledge retention was also problematic. Performance on the original Synapse task degraded substantially, with average DSC dropping from 0.795 to 0.683. Individual organs showed systematic decline: spleen DSC fell from 0.939 to 0.807, and pancreas from 0.755 to 0.631. This indicates that even selective parameter freezing could not prevent significant catastrophic forgetting.

TPGM Performance:

TPGM demonstrated superior results across both adaptation and retention metrics. On KiTS, it achieved remarkable performance with tumor DSC of 0.819 and kidney DSC of 0.936 - surpassing not only Surgical Fine-Tuning but also the from-scratch baseline. This suggests that TPGM's learnable regularization successfully leveraged pre-trained features while providing sufficient flexibility for specialization.

Knowledge retention was also improved, with average Synapse DSC maintained at 0.713 compared to 0.683 for Surgical Fine-Tuning. Individual organs showed better preservation: spleen DSC of 0.834 and pancreas DSC of 0.677. TPGM's adaptive, layer-wise constraints proved more effective at balancing plasticity and stability.

Multi-Stage Catastrophic Forgetting:

The three-task sequence (Synapse to KiTS to LiTS) using standard fine-tuning revealed complete catastrophic forgetting. While the final model performed well on LiTS (liver DSC: 0.914, tumor DSC: 0.479), it showed zero performance on KiTS and near-zero performance on Synapse. This demonstrates the fundamental limitation of unconstrained parameter updates in sequential learning scenarios.

These results establish clear performance benchmarks and validate the need for principled fine-tuning approaches. TPGM emerges as the strongest individual method, successfully improving both task adaptation and knowledge retention compared to standard approaches. The dramatic failure of sequential full fine-tuning provides strong motivation for the proposed hybrid framework that combines the theoretical foundations of Surgical Fine-Tuning with the adaptive capabilities of TPGM.

TODO: ADD SURGICAL AND TPGM RESULTS

TODO: ADD SURGICAL+TPGM MIX RESULTS AFTER FIXES

TODO: ADD TABLES/GRAPHS

TODO: ADD MORE DETAILS ABOUT TPGM/SURGICAL TUNING

VI. CONCLUSION AND FUTURE WORK

TODO: CONCLUSION

REFERENCES

1. Bilic, P. et al. (2023) 'The liver tumor segmentation benchmark (lits)', *Medical Image Analysis*, 84, p.102680.

2. Lee, Y., Chen, A.S., Tajwar, F., Kumar, A., Yao, H., Liang, P. and Finn, C. (2022) 'Surgical fine-tuning improves adaptation to distribution shifts', *arXiv preprint arXiv:2210.11466*.

3. Liu, X. et al. (2025) 'CSWin-UNet: Transformer UNet with cross-shaped windows for medical image segmentation', *Information Fusion*, 113, p.102634.

4. Myronenko, A., Yang, D., He, Y. and Xu, D. (2023) 'Automated 3D segmentation of kidneys and tumors in MICCAI KiTS 2023 challenge', in *International Challenge on Kidney and Kidney Tumor Segmentation*. Cham: Springer Nature Switzerland, pp. 1-7.

5. Tian, J., Dai, X., Ma, C.Y., He, Z., Liu, Y.C. and Kira, Z. (2023) 'Trainable Projected Gradient Method for Robust Fine-Tuning', in *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7836-7845.