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 [2] demonstrate remarkable accuracy by effectively capturing complex spatial hierarchies in medical scans. While training from scratch remains common in medical imaging, transfer learning - where models pre-trained on large-scale datasets are fine-tuned for downstream tasks - is becoming increasingly important with the rise of medical image foundation models. However, as these models grow in size and complexity, full fine-tuning demands extensive computational resources and time due to the massive number of operations involved, posing a significant barrier in data-limited and resource-constrained settings.

To address this challenge, various parameter-efficient fine-tuning methods have been developed [1,4]. While these techniques reduce training costs by updating only a subset of parameters (e.g., through manual layer freezing where only selected layers are trained), early approaches often relied on simple heuristics or introduced architectural modifications like adding new adapter modules. Recent works have shown more principled, automated approaches. For instance, Surgical Fine-Tuning [3] proposes identifying and preserving a core feature subspace to prevent catastrophic forgetting - the phenomenon where models lose previously learned knowledge when adapting to new tasks - while the Trainable Projected Gradient Method (TPGM) [4] 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. Modern architectures, such as CSWin-UNet, will be adapted and evaluated within this framework. The core objective is to develop an algorithm that automatically determines an optimal fine-tuning strategy, balancing high segmentation accuracy with computational efficiency.

A significant contribution of this project is the integration of a multi-task preservation mechanism - a technique that maintains performance on previously learned tasks while adapting to new ones. This approach 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 regularization techniques in methods like TPGM [4] and Surgical Fine-Tuning [1], the proposed method incorporates adaptive parameter selection to isolate task-specific updates. By dynamically identifying and preserving parameters critical to prior tasks while allowing selective modification of others, this mechanism enables the model to retain knowledge from previously learned tasks, effectively mitigating catastrophic forgetting and cross-task interference when 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.

1. CSWin-UNet for Efficient Medical Image Segmentation

CSWin-UNet [2] (Figure 1) was developed to address a fundamental challenge in medical image segmentation: capturing global context like Transformers do, but without their massive computational cost. Traditional CNN-based models like UNet [5] work well but struggle with long-range dependencies, while Transformer models can capture global information but are computationally expensive. Previous attempts like Swin-UNet [6] tried to bridge this gap with windowed attention but still had limited interactions between different parts of the image.

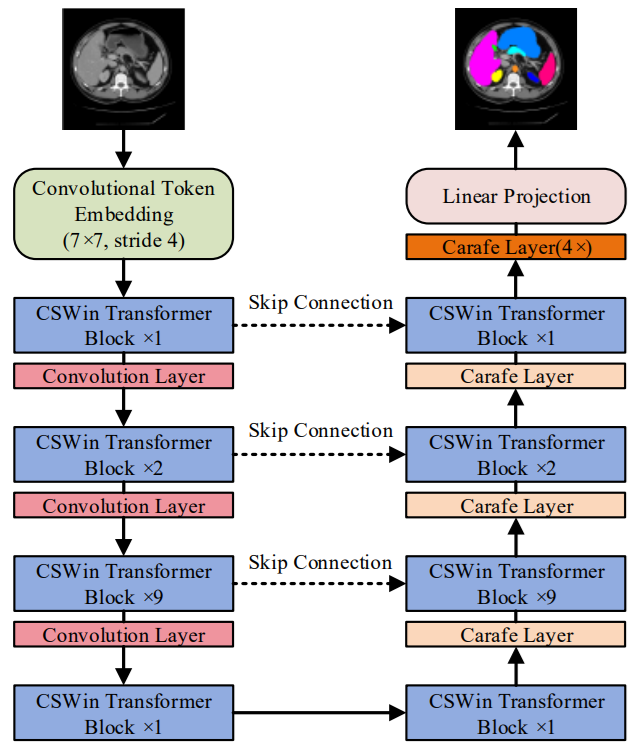
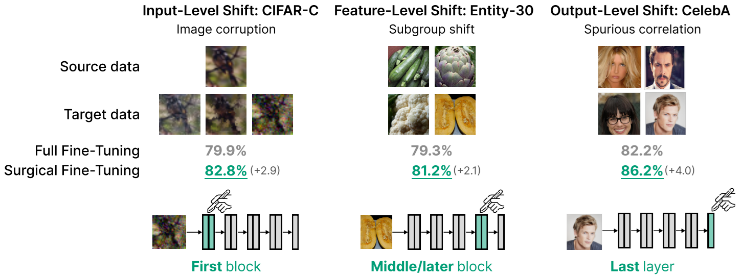
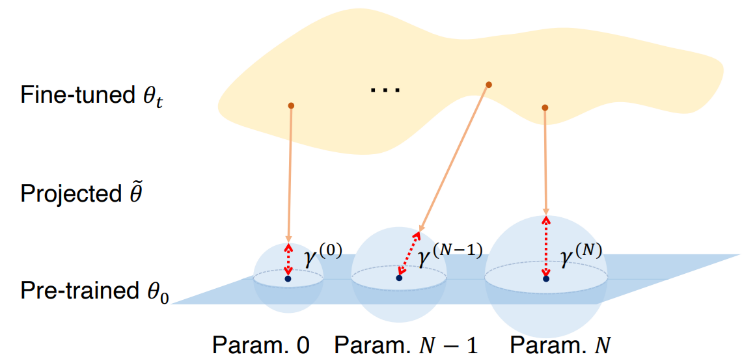
CSWin-UNet's main innovation is the cross-shaped window attention mechanism [2]. Instead of using square windows like previous methods, it processes information along horizontal and vertical strips simultaneously. This simple change dramatically expands the receptive field without significantly increasing computational cost. The model also employs CARAFE (Content-Aware ReAssembly of FEatures) [7], a content-aware upsampling method that dynamically predicts reassembly kernels based on feature content, resulting in sharper and more accurate segmentation boundaries.

Figure 1. CSWin-UNet architecture

On the Synapse multi-organ CT dataset, CSWin-UNet achieved 81.12% Dice similarity coefficient [2], outperforming both CNN-based methods like UNet [5] and other Transformer models like Swin-UNet [6]. More importantly for this project, it accomplished this with only 23.57M parameters - significantly fewer than typical Transformer architectures (Swin-UNet: 41.58M [6], ViT-Base: ~86M [8]) and more efficient than many CNN variants, making it one of the most parameter-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.

1. Preserving Core Features via Surgical Fine-Tuning

The core problem addressed by Lee et al. [1] is straightforward but critical: when adapting pre-trained models to new data distributions, standard fine-tuning methods can be suboptimal. While full fine-tuning provides flexibility, it risks overfitting on small target datasets and forgetting valuable knowledge acquired during pre-training. Conversely, linear probing (training only the final layer) is often too restrictive for effective adaptation, especially under significant distribution shifts.

To address this, Lee et al. introduce Surgical Fine-Tuning, which involves fine-tuning only a small, contiguous subset of a model's layers while freezing the rest. Their central and counter-intuitive finding is that the optimal subset of layers to tune depends directly on the type of distribution shift between the source and target data. For input-level shifts (e.g., image corruptions in CIFAR-C), they demonstrate that fine-tuning only the first few layers is most effective. For feature-level shifts (e.g., changes in sub-populations in Entity-30), tuning the middle layers works best. Finally, for output-level shifts (e.g., spurious correlations or flipped labels), tuning the last layer is superior, aligning with conventional wisdom. In all tested cases, the best surgical approach matched or significantly outperformed full fine-tuning.

Manually testing every possible subset of layers is impractical. Therefore, the authors investigate methods for automatically selecting which layers to tune. Their most successful approach, Auto-RGN, uses the Relative Gradient Norm (the ratio of a layer's gradient norm to its parameter norm) to assign a per-layer learning rate. The intuition is that layers with larger relative gradients are more informative for the new task and should be updated more. This automated method requires only a single fine-tuning run and was shown to consistently outperform full fine-tuning across seven different distribution shift benchmarks.

Figure 2. Types of distribution shifts

While the paper's framework (Figure 2) of categorizing distribution shifts as "input," "feature," or "output" level provides a powerful model, this is a high-level heuristic. Real-world shifts are often complex and multifaceted, making the ideal manual choice of which layers to tune non-obvious without extensive experimentation. This very challenge highlights the practical value of an automated method like Auto-RGN, which bypasses the need for a user to correctly diagnose the nature of the shift. A key trade-off remains, however: while Auto-RGN offers a practical, single-run solution, it has not yet achieved the peak performance of manually finding the optimal block via cross-validation, indicating that while its underlying heuristic is effective, it is not yet a perfect solution for layer selection.

1. 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.

Figure 3. TPGM weight projection

TPGM's key idea is to constrain each layer's weights to stay within a learnable "radius" of their original values (Figure 3). 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 insights into efficient model adaptation. CSWin-UNet [2] demonstrates that architectural efficiency is crucial for medical imaging applications, achieving strong performance with relatively few parameters. Surgical Fine-Tuning [1] and TPGM [4] 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, such as fine-tuning on related tasks over time (e.g., organ segmentation followed by tumor segmentation within that organ) or continuous model updates to prevent drift as new data streams in after deployment. 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 core idea of this work is to develop an automated fine-tuning framework that enhances the efficiency and robustness of medical image segmentation models by intelligently combining architectural efficiency with principled parameter adaptation. This section first defines the segmentation system mathematically and introduces the chosen base network architecture in Section 3.1. Subsequently, the two key components of the proposed hybrid fine-tuning approach - Surgical Fine-Tuning and the Trainable Projected Gradient Method - are detailed in Sections 3.2 and 3.3, respectively. Finally, Section 3.4 describes how these two methods are integrated to form the novel fine-tuning strategy.

3.1 Segmentation Network Architecture

For medical image segmentation, the task is to map an input medical image to a corresponding segmentation mask. Mathematically, given an input image (where H, W, and C represent height, width, and number of channels, respectively), the segmentation network aims to produce a predicted segmentation map (where K is the number of target classes or segments).

The base architecture for this research is CSWin-UNet [2], a Transformer model that has demonstrated state-of-the-art performance in medical image segmentation. This model was selected as the primary testbed because it effectively balances the global context modeling capabilities inherent to Transformers with a high degree of computational efficiency, making it particularly suitable for resource-constrained medical imaging applications.

CSWin-UNet addresses a fundamental limitation of standard Vision Transformers: their quadratic computational complexity, which often renders 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 model also employs CARAFE [7], a content-aware upsampling method that dynamically predicts reassembly kernels based on feature content, resulting in sharper and more accurate segmentation boundaries.

3.2 Surgical Fine-Tuning

The first component of the proposed hybrid approach builds on Surgical Fine-Tuning [1], which addresses a critical trade-off in transfer learning: while full fine-tuning offers flexibility, it risks catastrophic forgetting and overfitting, especially with small datasets. Conversely, linear probing (training only a new classifier on frozen features) often proves too restrictive for effective adaptation to new tasks. Surgical Fine-Tuning decomposes the learned feature space into "core" and "residual" subspaces, where core features represent essential, generalizable knowledge that should be preserved from the pre-trained model.

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.

3.3 Trainable Projected Gradient Method

The second component employs the Trainable Projected Gradient Method (TPGM) [4], 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.

3.4 Proposed Hybrid Fine-Tuning Method

The proposed method combines Surgical Fine-Tuning and TPGM into a coordinated optimization framework. Rather than applying these techniques sequentially, the hybrid approach creates a complementary relationship where surgical analysis dynamically adjusts learning rates while TPGM provides learnable regularization constraints to prevent excessive drift from pre-trained knowledge.

The surgical component operates at the beginning of each training epoch, performing automated gradient analysis on a small subset of training batches to maintain computational efficiency. Two gradient-based metrics are available for this analysis. The Relative Gradient Norm (RGN) method calculates weights based on the ratio of each layer's gradient norm to its parameter norm, effectively identifying layers where the new task requires the most significant relative updates. These weights are then used to create layer-specific learning rates, allowing more task-relevant layers to adapt faster while effectively reducing the learning rate for others.

The TPGM component operates on a separate schedule, running periodically (every N epochs after starting epoch M) using a dedicated subset of the training data. TPGM introduces learnable constraint parameters that define how far each layer's weights can deviate from their pre-trained values, using either L2 or MARS norm depending on configuration. These constraint radii are initialized to be permissive, particularly for final classification layers, and are optimized using a separate Adam optimizer. During TPGM phases, the method performs constraint optimization by minimizing the segmentation loss while simultaneously learning appropriate regularization boundaries for each layer.

The integration follows a coordinated schedule rather than nested optimization loops. Each training epoch begins with surgical gradient analysis to determine layer-specific learning rates, followed by standard gradient-based training using these adaptive rates. Periodically, TPGM constraint optimization is performed using its dedicated data subset to refine the regularization boundaries. The framework concludes with a final TPGM projection to ensure all parameters respect the learned constraints.

To address sequential learning challenges, the framework incorporates several specialized mechanisms. The model's output layer is dynamically expanded to accommodate new classes through a label remapping scheme that preserves the original class structure. Knowledge distillation from the frozen previous-stage model provides regularization to maintain performance on old tasks. A Focal Loss with inverse square root frequency-based class weighting addresses severe class imbalance in medical segmentation datasets. This comprehensive approach enables effective specialization on new tasks while maintaining robust knowledge retention from previous stages.

IV. EXPERIMENTAL SETUP

This section details the datasets, preprocessing steps, continual learning protocol, implementation specifics, and evaluation metrics used to assess the proposed method.

A. Datasets

To construct a clinically relevant sequential learning scenario, three publicly available abdominal CT datasets were selected: Synapse (multi-organ segmentation) [9], KiTS23 (kidney tumor segmentation) [10], and LiTS17 (liver tumor segmentation) [11]. This progression from general anatomy to specific pathologies is designed to mirror real-world clinical applications where foundational models are adapted for specialized diagnostic tasks.

The Synapse dataset serves as the foundational task, containing 30 CT scans with expert annotations for eight abdominal organs: aorta, gallbladder, left kidney, right kidney, liver, pancreas, spleen, and stomach. For the second task, the KiTS19 dataset was used, which focuses on kidney pathology with 210 cases providing annotations for kidneys and renal tumors. The third task utilizes the LiTS dataset, which targets liver pathology with 131 contrast-enhanced CT scans containing liver and liver tumor segmentations. To evaluate model robustness against common image quality degradation, a modified version of the Synapse dataset, termed Synapse-Blurred, was created by applying a Gaussian blur (σ=1.5) to each slice, simulating a domain shift.

B. Data Preprocessing and Splits

Prior to training, all CT volumes underwent a standardized preprocessing pipeline. Voxel intensities, expressed in Hounsfield Units (HU), were clipped to a window of [-1024, 1024] to remove irrelevant values and then normalized to a floating-point range of [0, 1]. All volumes were resampled to an isotropic spacing of 1.0 x 1.0 x 1.0 mm³ to ensure consistent anatomical scaling. For model input, 2D axial slices were extracted and resized to 224 x 224 pixels.

For each dataset, the data was partitioned at the patient level into training and testing sets. This separation ensures that no data from a single patient appears in more than one set, preventing data leakage and providing a more reliable estimate of generalization performance.

C. Continual Learning Protocol and Baselines

The evaluation follows a three-stage continual learning protocol designed to measure both task-specific performance and knowledge retention:

* Task 1: Train a segmentation model on the Synapse dataset (general abdominal anatomy).
* Task 2: Fine-tune the Task 1 model on the KiTS dataset (kidney tumors).
* Task 3: Fine-tune the Task 2 model on the LiTS dataset (liver tumors).

Each fine-tuning method under investigation (standard full fine-tuning, Surgical Fine-Tuning, TPGM, and the proposed hybrid approach) was evaluated through this sequence. To quantify catastrophic forgetting, the final model from Task 3 was tested on the test sets of all three tasks (Synapse, KiTS, and LiTS). Success is defined by strong performance on the current task (LiTS) while retaining high performance on previously learned tasks.

Two critical baselines were established for comparison: (1) an upper-bound baseline, where models were trained from scratch on each dataset independently, and (2) a lower-bound baseline using standard full fine-tuning, where all model parameters are updated during sequential training.

D. Implementation and Training Details

All experiments were implemented in Python using the PyTorch framework. Models were trained on an NVIDIA RTX 4060 Ti GPU. The AdamW optimizer was used with an initial learning rate of 0.001, a batch size of 32, and weight decay of 0.01. Models for each task were trained for 30 epochs using a cosine annealing learning rate scheduler. Model checkpoints were saved periodically every 5 epochs and a final model was saved at the end of training. The final trained model was used for evaluation on the test set.

E. Evaluation Metrics

Model performance was quantitatively assessed using two standard metrics for segmentation accuracy: the Dice Similarity Coefficient (DSC) and the 95th percentile Hausdorff Distance (HD95).

The Dice Similarity Coefficient measures the volumetric overlap between the predicted segmentation (P) and the ground truth segmentation (G). It is defined as:

where is the number of voxels in the intersection of the predicted and ground truth masks, and |P| and |G| are the total number of voxels in each mask, respectively. DSC values range from 0 (no overlap) to 1 (perfect overlap).

The Hausdorff Distance measures the maximum distance between the surfaces of the predicted and ground truth volumes, providing an assessment of boundary accuracy. To reduce sensitivity to outliers, the 95th percentile Hausdorff Distance (HD95) was used. It is defined as the 95th percentile of the distances between each point on one surface and the closest point on the other. A lower HD95 value, measured in millimeters, indicates a better alignment between the predicted and ground truth boundaries.

V. RESULTS AND ANALYSIS

5.1 Baseline Performance: Standard Fine-Tuning

Table 1a: Pre-training Performance on Individual Datasets

|  |  |  |
| --- | --- | --- |
| Dataset | Class | Pre-train Performance |
| KiTS23 | Kidney | 0.915 (4.8) |
|  | Tumor | 0.759 (8.3) |
|  | Cyst | 0.463 (25.0) |
| LiTS17 | Liver | 0.735 (16.4) |
|  | Tumor | 0.527 (48.9) |

The pre-training results establish strong baseline performance when models are trained individually on each dataset. KiTS23 pre-training achieved excellent kidney segmentation (0.915 Dice) with precise boundaries (4.8 HD95), while tumor and cyst segmentation showed progressively lower performance reflecting their increasing segmentation difficulty. LiTS17 pre-training demonstrated robust liver segmentation (0.735 Dice) but faced challenges with tumor detection (0.527 Dice), consistent with the inherent difficulty of liver lesion segmentation.

To establish a performance baseline and demonstrate catastrophic forgetting severity, experiments were conducted using standard full fine-tuning across the three-task sequential learning protocol. Table 1 shows the progressive performance degradation as the model learns new tasks without knowledge preservation mechanisms.

Table 1b: Sequential Fine-Tuning Performance Across All Tasks

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Class | After Synapse Training | After KiTS23 Fine-tuning | After LiTS17 Fine-tuning |
| Synapse | Aorta | 0.864 (11.1) | 0.746 (43.3) | 0.745 (42.1) |
|  | Gallbladder | 0.598 (30.0) | 0.031 (18.7) | 0.000 (0.0) |
|  | Left Kidney | 0.812 (53.3) | 0.692 (37.0) | 0.610 (32.6) |
|  | Right Kidney | 0.755 (54.9) | 0.530 (78.0) | 0.385 (47.2) |
|  | Liver | 0.939 (29.3) | 0.585 (60.9) | 0.632 (63.7) |
|  | Pancreas | 0.659 (13.3) | 0.003 (75.4) | 0.000 (32.7) |
|  | Spleen | 0.912 (13.8) | 0.248 (31.4) | 0.135 (44.3) |
|  | Stomach | 0.821 (19.1) | 0.400 (40.2) | 0.340 (43.0) |
| KiTS23 | Kidney | - | 0.711 (48.4) | 0.614 (56.0) |
|  | Tumor | - | 0.635 (55.8) | 0.551 (78.0) |
|  | Cyst | - | 0.401 (77.4) | 0.351 (86.6) |
| LiTS17 | Liver | - | - | 0.753 (46.0) |
|  | Tumor | - | - | 0.550 (64.0) |

The initial model trained on Synapse achieved strong performance across all eight anatomical structures, with particularly robust segmentation for organs with clear boundaries (liver: 0.939, spleen: 0.912). When subsequently fine-tuned on KiTS23, the model achieved moderate performance on the new task (kidney: 0.711, tumor: 0.635) but suffered severe catastrophic forgetting on Synapse, with gallbladder and pancreas segmentation collapsing almost entirely. After final fine-tuning on LiTS17, complete knowledge loss occurred for several tasks, with both gallbladder and pancreas achieving 0.000 Dice scores.

The analysis reveals that smaller anatomical structures with lower initial performance are more susceptible to complete knowledge loss, while organs sharing anatomical relationships with new targets show better preservation. HD95 measurements indicate that boundary accuracy degrades substantially even when volumetric overlap is partially preserved.

5.2 Surgical Fine-Tuning Performance

Table 2 presents a comprehensive comparison between standard and surgical fine-tuning approaches across all tasks and training stages. The surgical method aims to mitigate catastrophic forgetting by identifying and preserving critical feature directions while allowing adaptation in orthogonal feature spaces.

Table 2: Comparison of Standard Fine-tuning vs Surgical Fine-tuning Performance

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Class | Standard FT | Surgical FT |
| Synapse | Aorta | 0.745 (42.1) | 0.844 (10.1) |
|  | Gallbladder | 0.000 (0.0) | 0.607 (11.0) |
|  | Left Kidney | 0.610 (32.6) | 0.786 (52.8) |
|  | Right Kidney | 0.385 (47.2) | 0.713 (50.2) |
|  | Liver | 0.632 (63.7) | 0.912 (10.1) |
|  | Pancreas | 0.000 (32.7) | 0.607 (13.0) |
|  | Spleen | 0.135 (44.3) | 0.912 (8.5) |
|  | Stomach | 0.340 (43.0) | 0.817 (16.2) |
| KiTS23 | Kidney | 0.614 (56.0) | 0.620 (35.3) |
|  | Tumor | 0.551 (78.0) | 0.549 (90.7) |
|  | Cyst | 0.351 (86.6) | 0.403 (31.9) |
| LiTS17 | Liver | 0.753 (46.0) | 0.693 (52.0) |
|  | Tumor | 0.550 (64.0) | 0.565 (52.3) |

Surgical fine-tuning demonstrated substantial improvements across the majority of segmentation tasks. Most notably, it completely prevented catastrophic forgetting for vulnerable structures like gallbladder (0.607 vs 0.000) and pancreas (0.607 vs 0.000), while maintaining competitive performance on new tasks. The method achieved significant improvements for spleen, stomach, and kidneys, with Dice score improvements ranging from 0.176 to 0.777 compared to standard fine-tuning.

The slight performance decrease on the final LiTS17 liver segmentation task (0.693 vs 0.753 Dice) represents a reasonable trade-off for preserving knowledge across eight anatomical structures and three distinct segmentation challenges. This 6% decrease enabled maintenance of meaningful segmentation performance across all previous tasks, demonstrating the practical value for sequential learning in medical imaging.5.3 TPGM Fine-Tuning Performance

5.3 TPGM Fine-Tuning Performance

Table 3 demonstrates the TPGM approach's effectiveness in preserving knowledge while adapting to new tasks through adaptive layer-wise regularization. TPGM showed remarkable stability in maintaining performance levels very close to the original baseline across all organ classes when transitioning between tasks.

Table 3: TPGM vs Standard Fine-Tuning Performance Across Sequential Learning Protocol

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Class | Standard FT | TPGM FT |
| Synapse | Aorta | 0.745 (42.1) | 0.856 (12.1) |
|  | Gallbladder | 0.000 (0.0) | 0.633 (28.7) |
|  | Left Kidney | 0.610 (32.6) | 0.807 (56.9) |
|  | Right Kidney | 0.385 (47.2) | 0.718 (72.1) |
|  | Liver | 0.632 (63.7) | 0.903 (32.8) |
|  | Pancreas | 0.000 (32.7) | 0.664 (13.3) |
|  | Spleen | 0.135 (44.3) | 0.908 (12.7) |
|  | Stomach | 0.340 (43.0) | 0.815 (15.7) |
| KiTS23 | Kidney | 0.614 (56.0) | 0.701 (50.8) |
|  | Tumor | 0.551 (78.0) | 0.555 (101.7) |
|  | Cyst | 0.351 (86.6) | 0.401 (115.7) |
| LiTS17 | Liver | 0.753 (46.0) | 0.782 (43.0) |
|  | Tumor | 0.550 (64.0) | 0.579 (59.5) |

TPGM successfully preserved performance for challenging structures that proved vulnerable under standard fine-tuning. The gallbladder not only maintained performance but improved (0.633 vs initial 0.598), while pancreas remained stable (0.664 vs initial 0.659). The liver segmentation proved particularly stable, maintaining virtually identical performance (0.903 vs initial 0.939) even after two sequential fine-tuning stages.

While TPGM provided substantial protection against catastrophic forgetting, some gradual performance erosion became apparent under extended sequential learning scenarios. The right kidney experienced the most significant degradation (0.718 final vs 0.755 initial), suggesting that TPGM, while highly effective, is not completely immune to cumulative degradation across multiple adaptation stages.5.4 Mixed Surgical and TPGM Fine-tuning.

5.4 Hybrid Surgical and TPGM Fine-Tuning Performance

The hybrid approach combines surgical fine-tuning's precise feature direction preservation with TPGM's adaptive layer-wise regularization capabilities. Table 4 presents comprehensive results comparing all fine-tuning methods after complete sequential learning.

The method showed improvements over standard fine-tuning while attempting to balance the trade-offs between knowledge preservation and task adaptation. On the initial Synapse dataset, the approach achieved reasonable performance with liver segmentation at 0.906 and spleen segmentation at 0.898. The method successfully prevented catastrophic forgetting for challenging structures like gallbladder (0.603 vs 0.000 in standard FT) and pancreas (0.584 vs 0.000), though performance on these difficult classes remained moderate.

Table 4: Comprehensive Comparison of All Fine-Tuning Methods After Complete Sequential Learning

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Class | Standard FT | Hybrid FT |
| Synapse | Aorta | 0.745 (42.1) | 0.842 (12.0) |
|  | Gallbladder | 0.000 (0.0) | 0.603 (10.3) |
|  | Left Kidney | 0.610 (32.6) | 0.777 (54.3) |
|  | Right Kidney | 0.385 (47.2) | 0.697 (51.6) |
|  | Liver | 0.632 (63.7) | 0.906 (16.2) |
|  | Pancreas | 0.000 (32.7) | 0.584 (13.3) |
|  | Spleen | 0.135 (44.3) | 0.898 (8.9) |
|  | Stomach | 0.340 (43.0) | 0.812 (18.1) |
| KiTS23 | Kidney | 0.614 (56.0) | 0.719 (33.5) |
|  | Tumor | 0.551 (78.0) | 0.586 (101.9) |
|  | Cyst | 0.351 (86.6) | 0.410 (70.3) |
| LiTS17 | Liver | 0.753 (46.0) | 0.683 (49.2) |
|  | Tumor | 0.550 (64.0) | 0.569 (57.6) |

For the intermediate KiTS23 task, the hybrid method maintained adequate performance with kidney segmentation at 0.719 and tumor detection at 0.586. When compared to the pre-training baselines from Table 1a, the hybrid approach achieved 78.6% of the individual pre-training performance for kidney segmentation (0.719 vs 0.915) and 77.2% for tumor detection (0.586 vs 0.759). While these results represent improvements over standard fine-tuning, they indicate that the combined approach provides incremental rather than transformative benefits for intermediate task adaptation, with some performance degradation compared to task-specific pre-training.

On the final LiTS17 task, the hybrid approach achieved liver segmentation at 0.683, which represents 93.0% of the individual pre-training performance (0.735 from Table 1a) and outperformed the pre-training tumor detection baseline (0.569 vs 0.527). While better than standard fine-tuning, the liver performance still falls short of what TPGM alone achieved (0.782). This suggests that combining the two methods introduces some limitations in final task adaptation, potentially due to competing regularization strategies that may constrain the model's flexibility.

The analysis reveals that the hybrid approach offers a middle-ground solution with moderate improvements across most tasks. While it consistently outperforms standard fine-tuning and prevents complete catastrophic forgetting, it doesn't consistently match the best performance of either individual method or the pre-training baselines. The approach shows particular value in preventing complete knowledge loss (as seen with gallbladder and pancreas segmentation) but may sacrifice some of the specialized strengths that make TPGM or surgical fine-tuning effective in specific scenarios. From a computational efficiency perspective, the hybrid method offers significant advantages, requiring only approximately 1 hour per fine-tuning step compared to the 8 hours required for individual pre-training, making it practical for sequential learning scenarios despite some performance trade-offs.

The results suggest that the hybrid method provides a reasonable compromise for clinical applications where moderate performance across all tasks is preferred over optimizing for specific learning stages, particularly when computational resources and time constraints favor continual learning over repeated pre-training. However, the marginal improvements and performance gaps compared to individual pre-training indicate that practitioners might achieve better results by selecting TPGM or surgical fine-tuning based on their specific clinical workflow requirements rather than relying on the combined approach.

VI. CONCLUSION AND FUTURE WORK

This research introduced an automated layer selection framework for efficient fine-tuning of medical image segmentation models, specifically addressing the critical challenge of catastrophic forgetting in sequential learning scenarios common in clinical workflows. The study systematically evaluated and compared Surgical Fine-Tuning, the Trainable Projected Gradient Method (TPGM), and a novel hybrid approach using CSWin-UNet architecture across a three-stage sequential learning protocol. The experimental evaluation revealed that standard fine-tuning demonstrated severe catastrophic forgetting, with complete knowledge loss for vulnerable anatomical structures such as gallbladder and pancreas (Dice scores dropping to 0.000) and substantial degradation across most organs, while the proposed fine-tuning methods successfully mitigated this forgetting while maintaining competitive performance on new tasks.

The comparative analysis revealed distinct performance characteristics for each method. Surgical Fine-Tuning proved highly effective at preventing catastrophic forgetting, completely recovering performance for structures that suffered total knowledge loss under standard fine-tuning, though it showed a modest trade-off in final task performance with liver segmentation achieving 0.693 compared to 0.753 with standard fine-tuning. TPGM demonstrated the most balanced performance across all scenarios, providing excellent knowledge preservation while maintaining strong adaptation capabilities, achieving superior final task performance (liver: 0.782, tumor: 0.579) while simultaneously preserving performance on all previous tasks. The hybrid approach, while consistently outperforming standard fine-tuning and preventing complete catastrophic forgetting, did not consistently surpass either individual method, suggesting that the combination of techniques may introduce competing regularization strategies that constrain the model's flexibility.

This paper several contributions to medical image segmentation by providing systematic evaluation of automated layer selection methods in sequential learning scenarios, demonstrating that different anatomical structures exhibit varying vulnerability to catastrophic forgetting, and establishing evidence that TPGM offers superior balance between knowledge preservation and task adaptation. The practical framework enables deployment of sophisticated segmentation models in resource-constrained clinical environments where computational efficiency is paramount. However, several limitations should be acknowledged, including the study's scope being limited to three datasets within abdominal CT imaging, the hybrid approach showing diminishing returns compared to individual methods, and the focus on sequential rather than truly continual learning scenarios without thorough computational overhead analysis.

Future research should explore continuous learning scenarios where new data arrives incrementally rather than in discrete batches, more closely mimicking real clinical deployments, and investigate adaptive method selection that automatically chooses between different fine-tuning approaches based on task characteristics and available resources. Cross-modal adaptation extending the framework to other medical imaging modalities such as MRI, ultrasound, and X-ray represents a valuable direction, alongside clinical validation studies to assess real-world performance and quantify computational savings in practice. Advanced directions include uncertainty-guided adaptation incorporating confidence measures for intelligent parameter update decisions, meta-learning integration to learn optimal fine-tuning strategies from previous experiences, and anatomical relationship modeling that explicitly captures hierarchical and spatial relationships between organs to guide more intelligent preservation and adaptation strategies. This work establishes a practical foundation for deploying large-scale medical image segmentation models in clinical settings, with the demonstrated effectiveness of automated layer selection methods, particularly TPGM, providing a pathway toward more efficient and robust medical AI systems that can adapt to new tasks while preserving valuable previously learned knowledge.

REFERENCES

1. 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*.
2. Liu, X. et al. (2025) 'CSWin-UNet: Transformer UNet with cross-shaped windows for medical image segmentation', *Information Fusion*, 113, 102634.
3. 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.
4. 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.
5. Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Cham: Springer international publishing.
6. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2022, October). Swin-unet: Unet-like pure transformer for medical image segmentation. In European conference on computer vision (pp. 205-218). Cham: Springer Nature Switzerland.
7. Wang, J., Chen, K., Xu, R., Liu, Z., Loy, C. C., & Lin, D. (2019). Carafe: Content-aware reassembly of features. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 3007-3016).
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
9. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., & Klein, A. (2015). Multi-atlas labeling beyond the cranial vault. *URL: https://www. synapse. org*.
10. Heller, N., Wood, A., Isensee, F., Rädsch, T., Teipaul, R., Papanikolopoulos, N., & Weight, C. (Eds.). (2024). Kidney and Kidney Tumor Segmentation: MICCAI 2023 Challenge, KiTS 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, October 8, 2023, Proceedings (Vol. 14540). Springer Nature.
11. Bilic, P., Christ, P., Li, H. B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., ... & Menze, B. (2023). The liver tumor segmentation benchmark (lits). Medical image analysis, 84, 102680.