Continual Learning in 3D Vision: Mitigating Catastrophic Forgetting and Domain Shift via Deep Latent Adapters and Covariate-Guided Fusion

Paraskevi-Antonia Theofilou, Paraskevi Tzouveli, Stefanos Kollias
National Technical University of Athens, Greece
partheofilou@ails.ece.ntua.gr, tpar@mail.ntua.gr, stefanos@cs.ntua.gr

Abstract

Deploying deep learning for 3D medical image analysis across real-world clinical environments exposes two critical vulnerabilities: scanner-induced domain shift and severe clinical class imbalance. Standard fine-tuning across different hospital protocols triggers catastrophic forgetting, while traditional Mixture of Experts (MoE) architectures fail in autonomous settings, frequently leading to negative transfer during Out-of-Distribution (OOD) shifts. Furthermore, the extreme long-tailed nature of clinical screening induces majority-class collapse, entirely corrupting diagnostic sensitivity. To resolve these failures, we propose the Covariate-Guided Adapter Fusion (CGAF) framework, an autonomous, task-free routing system. We deploy full-rank, multi-layer Deep Latent Attention Adapters optimized via Supervised Contrastive Learning into a frozen base encoder, mathematically guaranteeing zero catastrophic forgetting of source domains. To prevent MoE negative transfer, the CGAF router utilizes margin-bounded prototypical cosine similarity and a gap-tolerance fallback, securely isolating expert noise during OOD events. Finally, we resolve majority-class collapse by natively integrating Prior-Aware Logit Adjustment and dynamic Youden's J optimal thresholding into the routing gate. Extensive experiments on a multi-hospital 3D CT scan benchmark demonstrate that CGAF consistently achieves state-of-the-art stability and plasticity across both Convolutional and Vision Transformer backbones. By mathematically mitigating catastrophic forgetting and actively rescuing minority-class sensitivity, our framework provides a robust architecture for continuous 3D vision deployment, establishing a highly secure, privacy-preserving foundation for clinical AI.

1. Introduction

A fundamental challenge in scaling medical AI is the inability of deep learning models to generalize reliably across unseen data distributions. In 3D medical imaging, even minor variations in scanner hardware, reconstruction algorithms, or acquisition protocols induce severe covariate shifts that actively degrade diagnostic performance [43]. Standard fine-tuning across disparate scanner calibrations triggers catastrophic forgetting [3], while clinical privacy regulations prohibit data-pooling methods like experience replay [39]. To bypass these restrictions, Parameter-Efficient Fine-Tuning (PEFT) and Mixture of Experts (MoE) architectures offer scalable solutions for Domain-Incremental Learning (Domain-IL) [18]. However, in autonomous clinical settings, these systems frequently fail. Softmax-based MoE routers lead to negative transfer during Out-of-Distribution (OOD) shifts by blindly blending incompetent experts [17], while the severe long-tailed imbalance of real-world medical datasets induces majority-class collapse [45].

To resolve these vulnerabilities, we propose the Covariate-Guided Adapter Fusion (CGAF) framework, engineered specifically for autonomous routing in highly imbalanced, multi-institutional environments. CGAF replaces fragile softmax gating with a mathematically bounded, prototype-driven decision mechanism. To ensure structural robustness, we explicitly decouple the system's defenses into two targeted interventions. First, to prevent negative transfer during OOD shifts, the router enforces strict margin-based confidence bounds to securely identify known domains, while simultaneously utilizing a gap-tolerance fallback and dynamically scaled top-k soft routing to manage uncertainty and isolate expert noise when encountering unfamiliar scans. Second, to combat the severe statistical bias inherent in long-tailed clinical datasets, we integrate a native posterior correction step that actively neutralizes majority-class collapse. Together, these mechanisms allow CGAF to perform highly reliable task-free inference across both Convolutional and Vision Transformer backbones without requiring manual task identities. In summary, the main contributions of our work are:

- Introduction of a novel parameter-isolation strategy using lightweight Deep Latent Adapters optimized via Focal and Supervised Contrastive Loss to eliminate catastrophic forgetting without violating clinical data privacy.

- Proposal of CGAF, a novel prototype-driven Sparse MoE router utilizing Margin-Based Hard Routing with Gap-Tolerance Noise Isolation and Top-K Soft Routing to enable reliable inference and prevent negative transfer.
- Resolution of majority-class bias via Prior-Aware Logit Adjustment and Youden's J Statistic, optimizing the classification boundary to rescue minority-class sensitivity.
- Extensive comparative evaluation of 3D CNN and Vision Transformer backbones under sequential domain shift, demonstrating how the proposed method stabilizes architectures against catastrophic majority-class collapse.

2. Related Work

2.1. Continual Learning and Parameter-Efficient Adaptation

Traditional Continual Learning (CL) methods mitigate catastrophic forgetting via weight regularization [22] or experience replay [35]. However, replay fundamentally violates strict clinical data privacy regulations [19] (e.g., GDPR). Parameter-Efficient Fine-Tuning (PEFT) [13, 26] offers a privacy-preserving alternative by training only marginal parameter subsets while permanently freezing the base model. Recent surveys [25] highlight that PEFT remains vastly underutilized in 3D medical vision, despite its proven ability to drastically reduce computational overhead and prevent catastrophic forgetting compared to full-network fine-tuning [8].

While standard PEFT architectures, such as Low-Rank Adaptation (LoRA) [14], provide baseline parameter isolation, they frequently lack the representational capacity required to handle complex 3D medical domain shifts. To address this gap, we deploy full-rank, domain-specific intermediate modules. By isolating target domain knowledge within these deep adapters, our pipeline mathematically guarantees zero source-domain forgetting without relying on privacy-violating replay buffers.

2.2. Representation Learning under Domain Shift

Scanner-induced domain shift is a primary bottleneck in clinical deployments. Unsupervised Domain Adaptation [9] violates CL constraints by requiring simultaneous access to source and target data. Test-Time Adaptation [40] avoids this by updating batch normalization statistics via entropy minimization; however, it frequently collapses toward the majority class prior under severe class imbalance [32].

To avoid the severe model collapse inherent to test-time optimization on imbalanced data, we instead rely on Representation Learning [4]. Rather than appending a single shallow module, we systematically inject Deep Latent Adapters with spatial gating [15] across multiple intermediate layers of the frozen base network to perform robust affine transformations. By sculpting the latent manifold using Supervised Contrastive Learning [21], we force features into tightly clustered, separable representations, establishing a robust foundation for downstream prototypical routing.

2.3. Mixture of Experts and Task-Free Inference

While Mixture of Experts (MoE) effectively scales model capacity [36], standard softmax gating often blindly blends predictions from incompetent experts. When deployed on OOD data, this injects mathematical noise, leading to negative transfer [1, 17, 41]. Furthermore, many Continual Learning (CL) systems impractically require a manual "Task-ID" during inference to select the correct weights.

We reimagine MoE as a Sparse Covariate-Guided Adapter Fusion (CGAF) system engineered for autonomous, Task-Free Inference. Utilizing non-parametric domain prototypes, we enforce strict margin-based hard routing for known distributions, while leveraging a gap-tolerance soft fallback and dynamically scaled top-k soft routing to securely manage OOD uncertainty. This mathematically isolates expert noise, prevents negative transfer, and completely eliminates the need for manual Task-IDs.

2.4. Long-Tailed Recognition

Clinical datasets are typically highly imbalanced, often containing over 90% healthy samples [28]. While techniques like Focal Loss [27] help the model learn better features, the final prediction often remains heavily biased toward the majority class. This leads to majority-class collapse, where the network simply predicts "healthy" for almost every sample. Architectures that lack built-in spatial assumptions, such as Vision Transformers (ViTs), are particularly vulnerable to this failure [20, 42].

This creates a dangerous clinical paradox: a model can achieve deceptively high overall accuracy while completely failing to detect the minority disease class (near-zero sensitivity). To resolve this, we structurally separate the dataset's statistical bias from the actual diagnostic features. By integrating Prior-Aware Logit Adjustment [30] and dynamic optimal thresholding via Youden's J Statistic [44] directly into the routing head, we mathematically correct the model's final probabilities. This active correction rescues models from majority-class collapse and successfully balances sensitivity and specificity.

3. Methodology

To address the sequential deployment of 3D medical vision models across heterogeneous hospital environments, we formulate the problem as Domain-IL. In this CL setting, the diagnostic task remains constant, but the input distribution undergoes severe covariate shift. Our methodology, as presented on Figure 1, guarantees the prevention of catastrophic forgetting while enabling autonomous adaptation through a unified pipeline. The training phase is directed

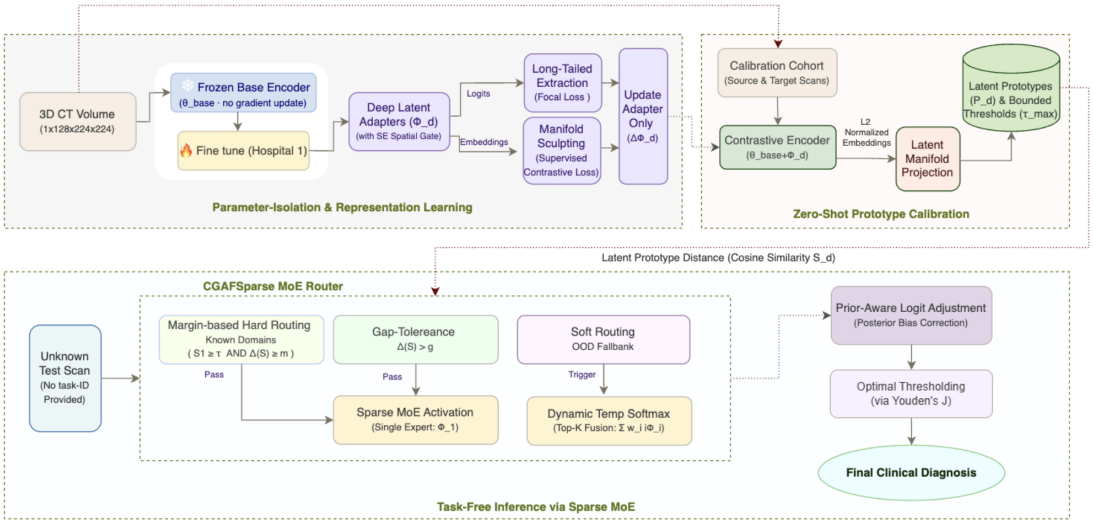

Figure 1. Pipeline diagram of the proposed CGAF methodology.

by (1) Parameter-Isolation via Deep Latent Adapters and (2) Representation Learning via Dual-Objective Optimization. During deployment, the system achieves autonomous operation through (3) Latent Prototype Calibration and (4) Task-Free Inference via Sparse MoE Routing (CGAF). Finally, to ensure diagnostic reliability under severe class imbalance, the network applies (5) Posterior Correction for Long-Tailed Recognition followed by (6) Optimal Clinical Thresholding via Youden's J Statistic.

3.1. Parameter-Isolation via Deep Latent Adapters

In standard CL paradigms [38], fully fine-tuning a network on a new target domain overwrites the weights required for the original source domain, causing catastrophic forgetting. To mathematically guarantee the preservation of prior knowledge, we employ a strict Parameter-Isolation strategy [29] using bottleneck adapters [13, 34]. We permanently freeze the foundational base model parameters, θ_{base}, and introduce domain-specific Deep Latent Attention Adapters (Φ_d). Crucially, rather than simply appending a shallow diagnostic head at the end of the network, these adapters are systematically injected directly into the deep intermediate layers of the architecture. Whether placed within the residual blocks of CNNs or the transformer blocks of ViTs, this intermediate integration allows the network to perform domain-specific affine transformations on high-level spatial and sequential features. Because the foundational weights remain strictly untouched, this approach prevents any corruption of the generalized feature extractor.

Let $x \in \mathbb{R}^{C\times D\times H\times W}$ represent an intermediate 3D spatial feature map, or $x \in \mathbb{R}^{B\times L\times C}$ for sequential tokens. The latent adapter first performs a bottleneck projection to limit the trainable parameter overhead:

$$z = \sigma(W_{down} \cdot \text{Norm}(x))$$

where $W_{down} \in \mathbb{R}^{C\times d_{bottle}}$ down-projects the channels to a lower-dimensional manifold ($d_{bottle} = 128$), $\text{Norm}(\cdot)$ applies domain-specific normalization and σ represents the GELU activation. The features are up-projected via $W_{up} \in \mathbb{R}^{d_{bottle}\times C}$ to yield the base adapted features $h_{adapt} = W_{up} \cdot z$.

To force the adapter to attend to sparse clinical anomalies, we integrate a Squeeze-and-Excitation (SE) spatial gate [15]. A channel descriptor $c = \text{GAP}(x)$ is generated via Global Average Pooling. The attention weights $g \in [0, 1]^C$ are computed as:

$$g = \delta(W_2 \cdot \sigma(W_1 \cdot c))$$

where W_1 reduces dimensionality by ratio r, W_2 restores it, and δ is the Sigmoid activation. The final residual output h' of the injected layer is the element-wise multiplication ($\odot$) of the adapted features and the attention gate, scaled by a learnable scalar s (initialized to 0.1):

$$h' = x + s \cdot \text{Dropout}(h_{adapt} \odot g)$$

3.2. Representation Learning via Dual-Objective Optimization

Standard Cross-Entropy (CE) loss [45] is insufficient for long-tailed continual learning tasks. CE establishes a linear decision boundary that frequently leads to majority-class collapse [31]. To enable our routing system, we frame the adapter optimization as a Representation Learning problem. The adapter's latent space must be topologically structured such that pathological features are tightly clustered and strictly separated from healthy prior distributions.

We optimize the domain-specific adapter parameters θ_{Φ_d} using a dual-objective loss function with $\lambda = 0.01$:

$$L_{total} = L_{Focal} + \lambda L_{SupCon}$$

First, Focal Loss [27] dynamically scales gradients based on prediction confidence, heavily penalizing the network for misclassifying rare minority volumes. For a predicted probability p_t, Focal Loss is defined as:

$$L_{Focal} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t)$$

where $\gamma = 2.0$ acts as the focusing parameter and α_t balances raw class frequencies.

Second, Supervised Contrastive Loss (L_{SupCon}) [21] is applied to the L_2-normalized deep embeddings, $v_i = f_{\theta_{base}, \Phi_d}(x_i)$. For a batch I, let $P(i)$ be the set of indices of all positive matches for an anchor i and $A(i)$ be all other indices. The loss is computed as:

$$\mathrm{L}_{SupCon} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(v_i \cdot v_p / \tau_{temp})}{\sum_{a \in A(i)} \exp(v_i \cdot v_a / \tau_{temp})}$$

where $\tau_{temp} = 0.07$. By forcing embeddings onto a unit hypersphere, L_{SupCon} creates the geometric separation required for our downstream Cosine Proximity router to function without a manual Task-ID. This mathematically forces intra-class embeddings to converge on the unit hypersphere, facilitating robust downstream clustering.

3.3. Latent Prototype Calibration for Zero-Shot Domain Classification

To achieve Task-Free Inference without catastrophic forgetting, the system must automatically identify a test scan's origin without needing a manual Task-ID or a separate domain classifier. We enable this zero-shot classification by using the non-parametric geometric center (the "prototype") of each hospital's contrastively trained feature space.

For a calibration set from domain d consisting of N_d samples, the domain prototype P_d is the L_2-normalized mean of the feature vectors:

$$P_d = \frac{\frac{1}{N_d} \sum_{i=1}^{N_d} v_i}{\left\| \frac{1}{N_d} \sum_{i=1}^{N_d} v_i \right\|_2}$$

To prevent OOD covariate shifts from triggering overly optimistic routing, we establish a strict similarity threshold τ_d based on the intra-domain variance. To account for latent space concentration, we bound this threshold to a theoretical clinical cap: $\tau_{d,\text{active}} = \min(\tau_d, 0.95)$

3.4. Task-Free Inference via Sparse MoE Routing - Covariate-Guided Adapter Fusion (CGAF)

Traditional MoE employs trainable softmax gating networks that either blindly average predictions from all experts via dense soft routing, which injects noise on OOD data, or activate only the highest-scoring modules via rigid top-k hard routing, which forces uncertain decisions. Our CGAF acts as a strict non-parametric Sparse MoE decision gate to prevent this negative transfer. It enforces margin-based hard routing via latent cosine proximity to securely isolate the optimal expert, while retaining a gap-tolerance fallback to our top-k soft routing for unfamiliar scans.

Given a test scan x_{test}, we compute the cosine similarity S_d against all loaded prototypes P_d:

$$S_d = \frac{f(x_{test}) \cdot P_d}{\|f(x_{test})\|_2 \|P_d\|_2}$$

The experts are sorted such that S_1 is the top match and S_2 is the runner-up. The final network prediction y is directed by a piecewise routing function utilizing a strict margin ($m = 0.05$) and a gap tolerance ($g = 0.20$):

$$y = \begin{cases} \Phi_1(x_{test}) & \text{if } S_1 \geq \min(\tau_1, 0.95) \wedge (S_1 - S_2) \geq m \\ \Phi_1(x_{test}) & \text{if } (S_1 - S_2) \geq g \\ \sum_{i=1}^{K} w_i \Phi_i(x_{test}) & \text{otherwise} \end{cases}$$

The first two conditions execute Hard Routing, isolating the single most competent expert to prevent noise injection. The third condition defaults to Top-K Soft Routing ($K = 2$). During Soft Routing, we apply Dynamic Temperature scaling ($T_{dyn} = T_{base} \cdot \max(S_1, 0.1)$) to the softmax function to ensure the marginally better expert retains greater influence. The fusion weights w_i are:

$$w_i = \frac{\exp(S_i / T_{dyn})}{\sum_{j=1}^{K} \exp(S_j / T_{dyn})}$$

Ultimately, this unified framework holistically benefits all deployments. By dynamically evaluating latent proximity and uncertainty, the system ensures that extreme clinical outliers, even those from known hospitals, are safely managed as OOD data rather than misrouted by a rigid gate.

3.5. Posterior Correction for Long-Tailed Recognition

Neural networks naturally learn the posterior probability $P(y|x) \propto P(x|y)P(y)$ according the Bayes' Theorem [5]. In long-tailed clinical datasets, the strong class prior $P(y)$ overwhelms the structural likelihood $P(x|y)$, causing architectures without strong spatial inductive biases to collapse into predicting the majority class. We mathematically decouple this dataset bias post-hoc.

Let L_{orig} denote the raw output logit from the adapter for a specific class y. Utilizing the known Bayesian prior π_y of the training dataset, we apply a Prior-Aware Logit Adjustment penalty [30]: $L_{adj} = L_{orig} - \beta \log(\pi_y)$, where $\beta = 1.0$. By subtracting the log-prior, we force the resulting distribution to rely entirely on the extracted anatomical features $P(x|y)$ rather than memorized dataset statistics.

3.6. Optimal Clinical Thresholding via Youden's J Statistic

The application of Logit Adjustment systematically shifts the output probability distribution, rendering the standard machine learning decision boundary of 0.5 clinically invalid. This misalignment causes a severe "seesaw effect," artificially inflating Sensitivity at the cost of Specificity.

To strictly optimize the expected clinical utility of the continual learning system, we dynamically calculate the optimal decision threshold over the Receiver Operating Characteristic (ROC) curve. Defining True Positive Rate (TPR) and False Positive Rate (FPR), the optimal threshold t^* is derived by maximizing Youden's J statistic [44]: $t^* = \arg\max_{t\in[0,1]} \Big(TPR(t) - FPR(t)\Big)$. Evaluating predictions at t^* establishes a mathematically rigorous equilibrium between diagnostic sensitivity and specificity, ensuring reliable anomaly detection across sequential hospital deployments.

4. Dataset

To evaluate the Covariate-Guided Adapter Fusion (CGAF) framework under realistic Domain-IL constraints, we utilize a dataset of 3D chest Computed Tomography (CT) scans across five distinct hospitals (H_1–H_5) [2, 23, 24]. Each volume is diagnostically annotated at the patient level as either COVID-19 positive or Normal (Healthy).

To enforce a strict continual learning environment, the datasets are partitioned by functional role. The Source Domain (H_1), as the most diverse set of volumetric samples, is used exclusively to train the foundational base encoder (θ_{base}), establishing a robust 3D feature extraction capability, which remains permanently frozen thereafter. The Target Domains (H_2, H_3, H_4, H_5) represent a sequential deployment scenario used solely to optimize the domain-specific Deep Latent Adapters (Φ_d), ensuring the acquisition of new clinical knowledge without catastrophic forgetting. Within this sequence, H_4 is designated as the primary ablation domain for architectural benchmarking.

Hospital ID	H_1	H_2	H_3	H_4	H_5	**Total**
Total Patients	1,769	2,103	494	2,866	298	7,530
Non-COVID	698	1,872	455	2,569	274	5,868
COVID-19	1,071	231	39	297	24	1,662

Table 1. Consolidated patient counts across all hospital domains.

Reflecting real-world clinical screening, this benchmark features a severe long-tailed class distribution as presented on Table 1. While H_1 exhibits moderate imbalance, the sequential target domains are heavily skewed. To prevent the adapters from collapsing into the majority-class prior during optimization, we apply standard 3D spatial augmentations at the preprocessing stage prior to dataloading, complemented architecturally by our Prior-Aware Logit Adjustment and Focal Loss mechanisms.

To standardize the heterogeneous scanner outputs into a unified latent space, raw CT slices for each patient are sequentially stacked into continuous 3D volumes. Spatial dimensions are resized to 224×224 via bilinear interpolation, and depth is standardized to $D = 128$ via trilinear interpolation, yielding a fixed input tensor of $1 \times 128 \times 224 \times 224$. Finally, voxel intensities are normalized to $[-1, 1]$ to stabilize gradient flow during contrastive representation learning.

5. Experiments

To evaluate our proposed pipeline, we conducted extensive experiments across multiple 3D architectures recommended for medical tasks. Our evaluation is designed to test both the parameter-efficient adaptation of the models and the autonomous routing capabilities of the system under multi-hospital severe domain shift and class imbalance.

5.1. Architectural Selection of the Latent Adapter

Before evaluating the multi-domain routing system, it was imperative to identify the most appropriate parameter-isolation architecture for extracting sparse pathological features under severe covariate shift. We designed a comprehensive adapter ablation study utilizing H_1 as the frozen source domain and H_4 as the target domain.

To evaluate plasticity and representation capacity, we established three baselines: Source-Only (Zero-Shot), Full Fine-Tuning (H_4) and a Target Upper Bound (H_4-Only). Keeping the base model (θ_{base}) frozen, we evaluated four distinct adapter configurations integrated into the intermediate layers: (1) a standard Bottleneck Adapter, (2) an Attention Adapter with a Squeeze-and-Excitation (SE) spatial gate, (3) a LoRA Attention Adapter with an SE gate and (4) our proposed Multi-Layer Deep Attention Adapter with Batch Normalization (BN) and Supervised Contrastive Loss (SupCon). Finally, to verify structural robustness, we performed component-level ablations on the best configuration, testing variants without the depthwise multi-layer structure, the SE attention gate and the domain-specific BN.

5.2. Task-Free Inference and Routing Robustness

Following the selection of the optimal parameter-isolation module, the foundational base model (θ_{base}) was pretrained on H_1 and frozen, with domain-specific adapters subsequently optimized for hospitals H_2 through H_5. To evaluate the CGAF Sparse MoE routing algorithm, we performed task-free inference across two deployment scenarios. In Scenario A (Known Domains), the router must autonomously identify the closest latent prototype among all four specialized adapters to execute the correct routing path utilizing Margin-Based Hard Routing. In Scenario B (OOD

Simulation), we conducted "Leave-One-Domain-Out" studies (e.g., Exclude H_4) by deleting a target domain's prototype and adapter from the system's memory. Under this OOD covariate shift, the strict margin and gap-tolerance thresholds fail, triggering the Dynamic Temperature Top-K Soft Routing fallback to infer a diagnosis.

To isolate the contributions of our routing components, we benchmarked the system against three baseline strategies: Naive Soft Routing (unconstrained ensembling), Entropy Soft Routing (weighting derived from entropy calculations) and Rigid Hard Routing (strict top-1 selection lacking fallback mechanisms). Furthermore, we conducted ablation studies to verify the impact of Prior-Aware Logit Adjustment, Youden's J thresholding and our proposed mathematical constraints, Margin, Gap Tolerance and Threshold Capping. By systematically removing these constraints to revert the system toward a naive ensembling state, we demonstrate their necessity in mitigating the uncertainty and class imbalance inherent in cross-hospital data.

5.3. Implementation Details and Hyperparameters

All experiments were implemented using PyTorch [33] and the MONAI [6] (Medical Open Network for AI) framework. The models were trained and evaluated on NVIDIA GPUs. We evaluated four distinct backbone architectures to observe differing inductive biases, ResNet-3D-18 [10, 12] and DenseNet-121 [16] for Convolution Neural Networks and Swin-UNETR [11, 37] and ViT-3D [7] for Vision Transformers. We also evaluate the performance of these models when either pre-trained with medical weights and fine-tuned on the source domain or trained from scratch.

The foundational base models were trained on H_1 and subsequently frozen in order to learn extracting features related with our task. For each target domain ($H_2 - H_5$), Deep Latent Adapters were injected into the intermediate layers and were trained for 20 epochs using a batch size of 8. We utilized the AdamW optimizer with an initial learning rate of 1×10^{-3}, modulated by a 'ReduceLROnPlateau' scheduler. To prevent overfitting, we employed an Early Stopping mechanism monitoring the validation loss.

To combat class imbalance during the adapter optimization, we utilized a 'WeightedRandomSampler' to enforce equal class representation within each mini-batch. The dual-objective loss was configured with a Focal Loss and a Supervised Contrastive Loss.

5.4. Evaluation Metrics

Due to the extreme long-tailed distribution of our clinical benchmark, standard Accuracy is highly deceptive, as it overwhelmingly reflects the majority-class prior. Therefore, we evaluate diagnostic viability strictly through balanced, threshold-independent and harmonic metrics: Area Under the Receiver Operating Characteristic (ROC-AUC), Area Under the Precision-Recall Curve (PR-AUC) and the Macro F1-Score. At the optimal Youden's J operating threshold, we also report Sensitivity (True Positive Rate) and Specificity (True Negative Rate). Clinically, high Sensitivity must be prioritized to prevent missed pathological diagnoses, while robust Specificity ensures the efficient minimization of false positive alarms.

Beyond standard classification, we evaluate the framework as a Domain-IL system by measuring the trade-off between knowledge acquisition and retention. We quantify Target Plasticity as the performance gain of domain-specific adapters (H_2–H_5) over the frozen zero-shot baseline. Simultaneously, we assess Source Stability by monitoring H_1 diagnostic performance throughout the sequence. This allows us to empirically verify our pipeline's mathematical guarantee of zero catastrophic forgetting, ensuring that the integration of new hospital data never degrades the original source knowledge.

6. Results

To validate the proposed CGAF framework, we structure our evaluation into three core phases: (1) an architectural ablation isolating the optimal parameter-efficient module, (2) a plasticity assessment of the model's domain-incremental learning capabilities across sequential domain shifts and (3) a robustness evaluation of the autonomous routing mechanism under OOD conditions.

6.1. Architectural Ablation of Latent Adapters

Table 2 presents an architectural ablation on the H_4 target domain to isolate the optimal parameter-efficient module for capturing complex 3D covariate shifts. Standard zero-shot inference using a frozen H_1 encoder generalizes poorly (e.g., 46.67% F1 for ViT-3D), while full fine-tuning establishes an empirical upper bound at the cost of catastrophic forgetting. Furthermore, standard parameter-efficient methods adapted from natural language processing, such as Low-Rank Adaptation (LoRA) and Simple Bottleneck Adapters, lack the representational capacity for 3D medical imaging. For instance, on ResNet-3D-18, LoRA achieves an 82.70% F1-score and 91.80% AUC, whereas our proposed Multi-Layer Deep Latent Attention Adapter significantly outperforms it with an 85.32% F1-score and 96.01% AUC. This demonstrates that our method effectively closes the gap to full fine-tuning while strictly preserving source weights.

Component-level ablation further justifies this architectural design, as removing Squeeze-and-Excitation (SE) spatial gating or Batch Normalization (BN) causes notable performance degradation across both CNNs and ViTs. We conclude that full-rank, attention-gated intermediate adapters, structured with supervised contrastive loss, provide the optimal balance of high representational capacity and strict parameter isolation necessary for handling 3D domain shifts.

Method / Configuration	ResNet-18-3D (%)				DenseNet-121 (%)				Swin-UNETR (%)				ViT-3D (%)			
	F1	BAcc	AUC	PR	F1	BAcc	AUC	PR	F1	BAcc	AUC	PR	F1	BAcc	AUC	PR
Baselines																
Zero-Shot (Encoder H_1)	76.81	76.59	90.08	67.78	65.27	62.94	71.40	35.25	54.50	54.19	64.10	19.63	46.67	49.62	64.09	19.60
Full Fine-Tuning (H_4)	88.94	91.50	96.40	87.65	86.41	84.89	95.05	77.65	70.33	70.17	81.04	43.99	63.17	60.18	77.79	36.03
Adapter Configurations																
1. Simple Adapter	46.86	50.00	88.37	67.36	55.85	68.32	78.76	39.44	20.97	52.74	56.39	13.04	12.74	51.01	54.92	18.76
2. Attention Adapter (w/ SE Gate)	75.13	85.17	92.25	73.78	68.51	70.53	74.71	37.33	58.61	57.78	68.22	24.08	54.01	53.81	67.92	26.87
3. LoRA Attention	82.70	81.37	91.80	76.28	60.10	74.56	81.22	42.28	60.77	68.92	70.11	25.08	48.67	50.82	64.61	20.67
4. Proposed Deep Latent Adapter	**85.32**	**87.34**	**96.01**	**82.28**	**62.42**	**79.40**	**84.12**	**48.82**	**72.33**	**78.70**	**83.60**	**40.93**	**57.06**	**64.18**	**69.35**	**21.15**
Component Robustness Ablation																
5. Proposed Adapter w/o SE Gate	84.74	87.91	95.02	78.34	51.23	73.25	82.79	46.28	74.50	69.87	66.97	60.24	51.64	59.14	61.80	18.57
6. Proposed Adapter w/o BN	76.43	69.81	90.93	74.57	65.24	76.72	82.77	41.70	46.86	50.00	62.44	17.86	57.71	56.15	62.46	25.11

Table 2. Architectural ablation of Deep Latent Adapters for target domain H_4.

Architecture	Method	Target H_2 (%)				Target H_3 (%)				Target H_4 (%)				Target H_5 (%)			
		F1	BAcc	AUC	PR	F1	BAcc	AUC	PR	F1	BAcc	AUC	PR	F1	BAcc	AUC	PR
ResNet-18-3D	Zero-Shot (H_1)	79.14	78.50	90.47	71.46	68.65	70.77	73.77	45.93	76.81	76.59	90.08	67.78	69.35	67.67	69.77	57.02
	Fine-Tuned	86.59	85.78	95.18	81.05	60.81	59.68	76.06	42.28	88.94	91.50	96.40	87.65	76.89	70.00	78.60	66.99
	Proposed	**81.22**	**87.94**	**94.24**	**79.93**	**25.12**	**58.45**	**76.58**	**61.20**	**85.32**	**87.34**	**96.01**	**82.28**	**76.89**	**70.00**	**76.74**	**58.30**
DenseNet-121	Zero-Shot (H_1)	59.17	56.99	64.75	33.97	55.65	63.03	74.30	42.17	65.27	62.94	71.40	35.25	59.09	57.67	59.07	32.98
	Fine-Tuned	77.67	79.84	87.40	72.72	46.26	47.89	82.92	26.92	86.41	84.89	95.05	77.65	78.69	70.00	65.12	48.75
	Proposed	**79.11**	**79.43**	**90.05**	**70.27**	**63.17**	**72.10**	**81.16**	**40.98**	**76.00**	**84.49**	**87.65**	**54.20**	**51.05**	**69.07**	**77.21**	**47.70**
Swin-UNETR	Zero-Shot (H_1)	59.48	61.13	64.44	25.70	74.22	72.89	74.12	54.46	54.50	54.19	64.10	19.63	46.67	48.84	68.84	24.32
	Fine-Tuned	73.62	72.84	85.73	61.15	70.66	67.34	69.01	50.53	70.33	70.17	81.04	43.99	74.33	76.51	85.12	67.11
	Proposed	**60.85**	**71.68**	**77.94**	**32.50**	**31.02**	**41.90**	**59.15**	**55.02**	**72.33**	**78.70**	**83.60**	**40.93**	**77.67**	**77.67**	**80.93**	**65.56**
ViT-3D	Zero-Shot (H_1)	49.34	51.11	66.59	24.02	69.79	75.62	80.81	48.68	46.67	49.62	64.09	19.60	64.44	60.00	73.49	37.69
	Fine-Tuned	49.24	51.11	66.48	20.14	62.28	60.39	74.47	28.13	63.17	60.18	77.79	36.03	59.09	57.67	86.05	49.82
	Proposed	**52.25**	**52.68**	**61.98**	**24.35**	**64.36**	**72.80**	**68.31**	**41.52**	**57.06**	**64.18**	**69.35**	**21.15**	**44.83**	**45.35**	**52.09**	**12.62**

Table 3. Continual learning generalization and plasticity evaluation across sequential target domains.

6.2. Continual Learning Generalization Across Target Domains

Having established the optimal adapter architecture, Table 3 evaluates the framework's plasticity across a rigorous sequential deployment over four distinct hospital domains (H_2–H_5). The data reveals that scanner-induced domain shifts severely degrade the frozen base encoder's performance. For instance, DenseNet-121's zero-shot PR-AUC drops to 33.97% on H_2, demonstrating that high-level feature extraction cannot blindly generalize across disparate hospital protocols without targeted adaptation.

By injecting the domain-specific Proposed Multi-Layer Deep Attention Latent Adapters, the framework demonstrates exceptional model plasticity. Across all evaluated CNN and ViT architectures, the adapters successfully capture target-specific geometries, consistently and dramatically elevating diagnostic metrics above the zero-shot baseline. For example, on Swin-UNETR (H_4), the adapter improves the F1-score from 54.50% to 72.33%. Crucially, because these adapters operate strictly via parameter isolation, this robust target plasticity is achieved alongside absolute model stability. The source domain (H_1) weights are never corrupted, mathematically guaranteeing zero catastrophic forgetting across diverse scanner distributions.

6.3. Autonomous Routing and OOD Robustness

Table 4 provides a comparative evaluation of the proposed routing framework against baseline MoE strategies across both internal and out-of-distribution deployment scenarios, including simulated domain-exclusion tests, for H_4. In the "Known Domains" scenario, standard Naive Soft Ensembling and Entropy Routing suffer from severe majority-class collapse, particularly on architectures lacking spatial inductive biases. For instance, on Swin-UNETR, these baselines fail to detect a single pathological case on H_1, yielding 0.00% Sensitivity and defaulting to the 39.78% majority-class prior. Conversely, CGAF actively rescues the posterior distribution via Prior-Aware Logit Adjustment and dynamic thresholding. By mathematically decoupling dataset bias from diagnostic likelihood, CGAF dramatically elevates minority-class sensitivity, improving Swin-UNETR's source stability from 0.00% to 66.67% and target plasticity to 83.02%, while ViT-3D's source sensitivity jumps from 0.60% to 67.26% and target one to 59.06%.

The "OOD Simulation (Exclude H_4)" scenario further exposes the fragility of standard MoE routing. When encountering unfamiliar H_4 scans, standard routers lead to uncertainty-driven blending of incompetent experts, resulting in severe negative transfer (e.g., ResNet-3D-18 under

Architecture	Routing Method	"Known Domains" (Include $H_1 ... H_5$)								"OOD Simulation" (Exclude H_4)							
		Source Stability (H_1) (%)				Target Plasticity (H_4) (%)				Source Stability (H_1) (%)				Fallback Plasticity (H_4) (%)			
		Acc	AUC	Sens	Spec	Acc	AUC	Sens	Spec	Acc	AUC	Sens	Spec	Acc	AUC	Sens	Spec
ResNet-3D-18	Base Hard	76.34	98.21	60.71	100.0	88.39	96.11	1.89	100.0	76.34	98.21	60.71	100.0	N/A	N/A	N/A	N/A
	Naive Soft	44.44	97.14	7.74	100.0	88.62	91.14	3.77	100.0	49.82	96.99	16.67	100.0	89.29	88.92	9.43	100.0
	Entropy Soft	51.25	95.78	19.05	100.0	89.29	89.03	9.43	100.0	51.61	93.63	19.64	100.0	89.29	86.68	9.43	100.0
	Proposed CGAF	**87.81**	**94.11**	**93.45**	**79.28**	**73.66**	**81.54**	**73.58**	**73.67**	**77.42**	**89.09**	**66.07**	**94.59**	**72.54**	**77.34**	**69.81**	**72.91**
DenseNet-121	Base Hard	65.95	90.26	100.0	14.41	83.48	88.05	81.13	83.80	65.95	90.26	100.0	14.41	N/A	N/A	N/A	N/A
	Naive Soft	77.78	89.66	92.86	54.95	66.96	83.72	84.91	64.56	76.34	90.61	96.43	45.95	54.24	77.65	84.91	50.13
	Entropy Soft	79.57	88.58	89.29	64.86	68.08	81.95	81.13	66.33	76.34	90.59	94.05	49.55	57.81	77.80	83.02	54.43
	Proposed CGAF	**80.65**	**87.31**	**80.95**	**80.18**	**77.01**	**75.86**	**78.49**	**79.49**	**80.65**	**87.25**	**80.95**	**80.18**	**82.37**	**70.86**	**65.28**	**87.34**
Swin-UNETR	Base Hard	39.78	41.38	0.00	100.0	94.87	90.01	69.81	98.23	42.29	46.20	27.98	63.96	N/A	N/A	N/A	N/A
	Naive Soft	39.78	41.38	0.00	100.0	89.96	89.29	20.75	99.24	42.29	46.20	27.98	63.96	88.17	53.17	0.00	100.0
	Entropy Soft	39.78	41.38	0.00	100.0	91.29	87.01	37.74	98.48	42.29	46.20	27.98	63.96	88.17	53.17	0.00	100.0
	Proposed CGAF	**73.48**	**80.86**	**66.67**	**83.78**	**81.92**	**86.38**	**83.02**	**81.77**	**77.78**	**84.83**	**76.79**	**79.28**	**78.79**	**67.65**	**56.60**	**79.49**
ViT-3D	Base Hard	39.78	63.08	1.19	100.0	88.17	63.55	1.89	99.75	40.68	63.08	1.19	100.0	N/A	N/A	N/A	N/A
	Naive Soft	40.14	67.71	0.60	100.0	88.62	63.09	3.77	100.0	40.14	71.27	0.60	100.0	88.84	60.84	5.66	100.0
	Entropy Soft	40.14	62.69	0.60	100.0	88.17	63.37	0.00	100.0	40.14	71.27	0.60	100.0	88.17	61.09	0.00	100.0
	Proposed CGAF	**65.59**	**62.58**	**67.26**	**63.06**	**60.49**	**53.89**	**59.06**	**62.03**	**65.59**	**62.58**	**67.26**	**63.06**	**60.49**	**53.89**	**59.06**	**62.03**

Table 4. Robustness evaluation of standard MoE routing methods against our CGAF proposed system.

Entropy Routing manages only 9.43% Sensitivity). CGAF effectively mitigates this vulnerability through a dual-mechanism approach. For known domains, our margin-based hard routing securely leads to the correct expert, consistently matching or exceeding the theoretical upper bound for stability. Simultaneously, for unseen domains, CGAF's Gap-Tolerance Soft Fallback robustly isolates expert noise to provide vital OOD fallback plasticity (yielding 77.34% AUC and 69.81% Sensitivity for ResNet-3D-18), a capability structurally impossible for traditional hard routing. Ultimately, by enforcing these mathematically bounded routing gates, CGAF transforms MoE architectures into highly reliable, task-free continual learning systems suitable for autonomous clinical deployment.

6.4. Methodological Analysis

Our analysis demonstrates that CNNs naturally resist domain shifts better than ViTs due to their built-in spatial biases. Because ViTs lack these biases, they are highly susceptible to dataset imbalances. Under standard routing, they often collapse into predicting only the majority class, completely failing to detect minority cases (yielding 0.00% Sensitivity). However, applying CGAF's Prior-Aware Logit Adjustment neutralizes this vulnerability. This allows the system to leverage the high representational capacity of ViTs, significantly improving rare-class detection.

Traditional approaches struggle with the demands of multi-institutional 3D medical imaging. Zero-shot inference degrades heavily when encountering new scanners, while full fine-tuning catastrophically overwrites previously learned knowledge. Standard parameter-efficient methods like LoRA lack the depth required to process complex CT textural shifts. Even in multi-expert setups, standard MoE architectures fail during OOD shifts. Standard soft routing blindly averages predictions, which injects mathematical noise from unqualified experts and causes severe negative transfer. Conversely, standard hard routing rigidly forces a decision from a single expert, completely failing to manage the high uncertainty of unfamiliar clinical data. By replacing these flawed gates with our mathematically bounded, margin-driven router and gap-tolerance fallback, our CGAF framework directly overcomes these structural limitations.

7. Conclusion

In this paper, we introduced the Covariate-Guided Adapter Fusion (CGAF) framework to systematically resolve catastrophic forgetting and negative transfer in 3D medical imaging. By optimizing multi-layer domain-specific adapters integrated into a permanently frozen base model, CGAF mathematically guarantees that source knowledge is fully preserved as new hospital domains are acquired. To overcome the severe vulnerabilities of standard MoE routing, which frequently collapses on imbalanced or unfamiliar clinical data, our advanced autonomous routing system securely isolates expert noise during OOD shifts and actively rescues minority-class sensitivity. Ultimately, CGAF delivers a highly scalable, privacy-preserving foundation for continuous, task-free clinical deployment across hospitals without centralizing sensitive patient data. Future work will integrate Vision-Language Models (VLMs) and Retrieval-Augmented Generation (RAG) into the CGAF architecture. Synthesizing 3D scans with retrieved anonymized clinical text will guide routing decisions and ensure robust multi-modal reasoning for highly unfamiliar cases.

Acknowledgements

This work was supported by computational time granted from the National Infrastructures for Research and Technology S.A. (GRNET) in the National HPC facility - ARIS.

References

[1] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3366–3375, 2017. 2

[2] Anastasios Arsenos, Dimitrios Kollias, and Stefanos Kollias. A large imaging database and novel deep neural architecture for covid-19 diagnosis. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pages 1–5, 2022. 5

[3] Nourhan Bayasi, Ghassan Hamarneh, and Rafeef Garbi. GC2: Generalizable continual classification of medical images. *IEEE Transactions on Medical Imaging (TMI)*, 43(11): 3767–3779, 2024. 1

[4] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. 2

[5] Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006. 4

[6] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Isaac Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022. 6

[7] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[8] Raman Dutt, Linus Ericsson, Pedro Sanchez, Sotirios A. Tsaftaris, and Timothy Hospedales. Parameter-efficient fine-tuning for medical image analysis: The missed opportunity. In *Proceedings of The 7th International Conference on Medical Imaging with Deep Learning (MIDL)*, pages 406–425. PMLR, 2024. 2

[9] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research (JMLR)*, 17(1):2096–2030, 2016. 2

[10] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 6

[11] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022*, pages 272–284, 2022. 6

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[13] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning (ICML)*, pages 2790–2799. PMLR, 2019. 2, 3

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. 2

[15] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018. 2, 3

[16] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 6

[17] Yinuo Jiang, Yan Xiaodong, Keyan Ding, Deng Zhao, Lei Liang, Qiang Zhang, and Huajun Chen. HiMoLE: Towards OOD-robust LoRA via hierarchical mixture of experts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 1, 2

[18] Yunzhi Zhuge Lu Zhang Ping Hu Dong Wang Huchuan Lu You He Jiazuo Yu, Zichen Huang. MoE-Adapters++: Toward more efficient continual learning of vision-language models via dynamic mixture-of-experts adapters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1

[19] Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311, 2020. 2

[20] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations (ICLR)*, 2020. 2

[21] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 18661–18673, 2020. 2, 4

[22] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. 2

[23] Dimitrios Kollias, Anastasios Arsenos, and Stefanos Kollias. A deep neural architecture for harmonizing 3-d input data analysis and decision making in medical imaging. *Neurocomputing*, 542:126244, 2023. 5

[24] Dimitrios Kollias, Anastasios Arsenos, and Stefanos Kollias. Domain adaptation, explainability fairness in ai for medical image analysis: Diagnosis of covid-19 based on 3-d chest ct-scans. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4907–4914, 2024. 5

[25] Pratibha Kumari, Joohi Chauhan, Afshin Bozorgpour, Boqiang Huang, Reza Azad, and Dorit Merhof. Continual

learning in medical image analysis: A comprehensive review of recent advancements and future prospects. *arXiv preprint arXiv:2402.04944*, 2024. 2

[26] Vladislav Lialin, Vijeta Deshpande, and Anna Rumshisky. Scaling down to scale up: A guide to parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.15647*, 2023. 2

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 2, 4

[28] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen AWM van der Laak, Bram van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017. 2

[29] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7765–7773, 2018. 3

[30] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations (ICLR)*, 2021. 2, 4

[31] Kei Nitta, Taishi Nammoto, Kosuke Arima, and Toshikazu Ebisuzaki. Leveraging inductive bias in vit for medical image diagnosis. In *British Machine Vision Conference (BMVC)*, 2024. 3

[32] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Junzhou Huang. Sar: Stable test-time adaptation for open-world streams. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*. 2019. 6

[34] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3

[35] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2001–2010, 2017. 2

[36] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017. 2

[37] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20708, 2022. 6

[38] Gido M van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019. 3

[39] Tanvi Verma, Liyuan Jin, Jun Zhou, Jia Huang, Mingrui Tan, Benjamin Chen Ming Choong, Ting Fang Tan, Fei Gao, Xinxing Xu, Daniel S. Ting, and Yong Liu. Privacy-preserving continual learning methods for medical image classification: a comparative analysis. *Frontiers in Medicine*, 10:1227515, 2023. 1

[40] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021. 2

[41] Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11293–11302, 2019. 2

[42] Lu Yang, He Jiang, Qing Song, and Jun Guo. A survey on long-tailed visual recognition. *International Journal of Computer Vision*, 130(7):1837–1872, 2022. 2

[43] Haoran Ying, Yijun Lia, and Zedong Fu. A survey of generalization and adaptation in medical imaging foundation models. *Preprints.org*, 2025. 1

[44] William J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950. 2, 5

[45] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10795–10816, 2023. 1, 3