X-Edit: Exact, Explicit, and Explainable Null-Space Editing for Medical Vision Transformers

Yuanye Liu[1†], Siyuan Zhou[1†], Ke Zhang[2], Lei Li[3], Wei Chen[4], and Xiahai Zhuang[1✉]

[1] Fudan University, Shanghai, China
[2] Johns Hopkins University, Baltimore, USA
[3] National University of Singapore, Singapore
[4] University of Sydney, Sydney, Australia

Abstract. Pre-trained Vision Transformers (ViTs) are increasingly deployed for medical image classification. However, correcting their inevitable failure cases in dynamic clinical scenarios poses a critical challenge. Conventional fine-tuning approaches inherently suffer from catastrophic forgetting, severely degrading previously acquired diagnostic capabilities. Such instability fundamentally compromises clinical safety. Addressing this vulnerability requires an active, controllable, and reliable intervention mechanism that is both theoretically grounded and inherently interpretable. To this end, we propose **X-Edit** (**eX**act, **eX**plicit, and **eX**plainable Editing), an efficient null-space model editing framework. X-Edit transitions the editing process from iterative gradient-based optimization to a theoretically grounded, closed-form solution. Specifically, we first explicitly localize the influential layers via causal tracing governing the erroneous prediction. Subsequently, we construct an orthogonal null-space projection matrix from a curated anchor set. By geometrically constraining the exact parameter update strictly within this null space, we provide mathematical guarantees that the intervention rectifies targeted errors without perturbing established diagnostic representations. Extensive evaluations on six medical imaging benchmarks demonstrate that X-Edit comprehensively suppresses catastrophic forgetting while achieving superior edit success rates. Our code is available at `https://github.com/HenryLau7/X-Edit`.

Keywords: Model Editing · Trustworthy AI · Post-hoc Intervention · Catastrophic Forgetting

1 Introduction

Pre-trained Vision Transformers (ViTs) [3] have demonstrated remarkable efficacy across a wide range of medical image classification tasks [1, 7]. However,

† These two authors contributed equally.
✉ Corresponding authors: Xiahai Zhuang (zxh@fudan.edu.cn)

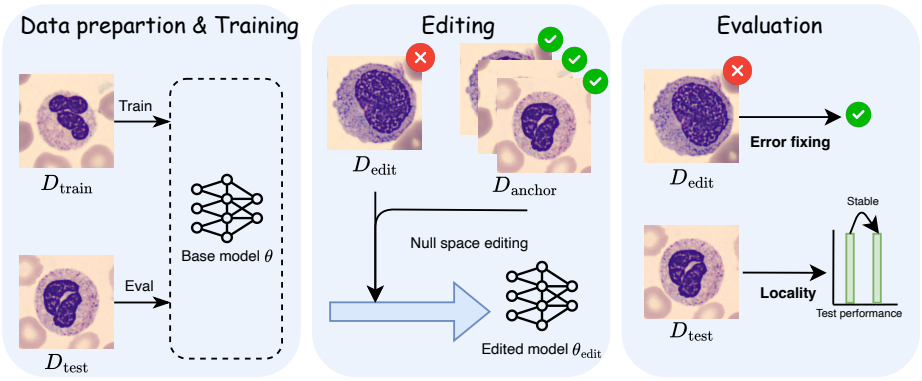

Fig. 1. Pipeline of X-Edit. A base model trained on D_{train} is edited to correct new error samples D_{edit} using null-space updates constrained by anchor samples D_{anchor}, preserving existing representations. The edited model is then evaluated for error fixing and locality on D_{edit} and a disjoint test set D_{test} to ensure stable performance.

when deployed in dynamic clinical environments, these models inevitably encounter failure cases. These errors often arise from online data streams containing long-tail pathological variants or domain shifts caused by different imaging equipment [5,13]. Continually ensuring the reliability of these models is a critical clinical imperative. The conventional approach of fine-tuning (FT) the model on these newly encountered error samples typically suffers from catastrophic forgetting, wherein the model loses its previously acquired diagnostic capabilities on correct samples. In clinical AI, such a phenomenon compromises clinical safety, as a model that forgets common medical knowledge after learning several new cases could not be trusted. Therefore correcting these errors requires an active, controllable, and reliable intervention that explicitly rectifies targeted failures without collateral damage [11].

To address this bottleneck, the concept of *Model Editing* has emerged as a promising post-hoc paradigm [6]. Recent works have attempted to adapt model editing to Vision Transformers. LWE [18] proposed a meta-learning approach that trains a hypernetwork to generate sparse binary masks, identifying a subset of parameters to be fine-tuned on failure cases. Concurrently, RefineViT [9] intervened on specific attention heads alongside a newly learned representation projection. However, these approaches often necessitate computationally intensive optimization pipelines and rely on iterative, gradient-based stochastic refinement. Crucially, they remain fundamentally misaligned with the strict reliability and interpretability required in clinical AI [2, 16], where they lack the theoretical grounding required to provide mathematical guarantees against performance degradation.

Null-space projection is a geometric constraint originally pioneered in continual learning [19] to strictly isolate parameter updates. This fundamental concept

has been successfully adapted to mitigate catastrophic forgetting in Large Language Models (LLMs) [4, 15], but its potential for ViTs remains unexplored.

To bridge this gap and fulfill the urgent clinical demand for trustworthy AI, we propose **X-Edit** (**eX**act, **eX**plicit, and **eX**plainable Editing), an efficient and scalable null-space model editing framework. As illustrated in Fig. 1, we establish a rigorous post-hoc intervention setting. Given a base model pre-trained on a standard training set, the editing phase actively rectifies newly encountered error samples (D_{edit}) by executing null-space updates constrained by a curated set of anchor samples (D_{anchor}). This design guarantees that the intervention achieves targeted corrections without perturbing previously learned representations. Finally, the framework includes a comprehensive evaluation phase, assessing the updated model on both the specific edit samples to verify error fixing and a strictly disjoint test set to guarantee diagnostic stability. Our main contributions are summarized as follows:

1. We introduce the null-space model editing paradigm into medical image analysis for the first time as far as we know. By reframing the catastrophic forgetting problem as a geometric constraint optimization, we provide a mathematically controllable and safe method to edit ViT without compromising previously learned diagnostic features.
2. We propose X-Edit, a comprehensive pipeline that seamlessly integrates mechanistic interpretability with an exact, closed-form analytical solution for parameter updating. This transparent mathematical design makes our framework highly scalable and exceptionally suited for long-term clinical deployment.
3. We comprehensively validate our framework across six diverse medical imaging datasets. Extensive evaluations demonstrate that X-Edit successfully translates its theoretical guarantees into empirical reliability, achieving a superior balance between edit success rate and overall test stability compared to existing baselines.

2 Methodology

2.1 Problem Formulation

Let f_θ denote a pre-trained ViT optimized for a specific medical image classification task, initially trained on a base dataset D_{train}. We define an unseen test set $D_{test} = \{(x_i, y_i)\}_{i=1}^{N}$ purely for evaluation, which strictly does not participate in any training or editing phases.

During deployment, the model may encounter challenging cases, yielding incorrect predictions. We define these misclassified instances as the edit set $D_{edit} = \{(x_e, y_e)\}_{e=1}^{M}$, where $f_\theta(x_e) \neq y_e$. The goal of model editing is to find a parameter perturbation $\Delta\theta$ such that the updated model $f_{\theta+\Delta\theta}$ correctly classifies D_{edit}. To prevent catastrophic forgetting, we introduce an anchor set $D_{anchor} = \{(x_a, y_a)\}_{a=1}^{K}$ consisting of accurately predicted samples from D_{train}.

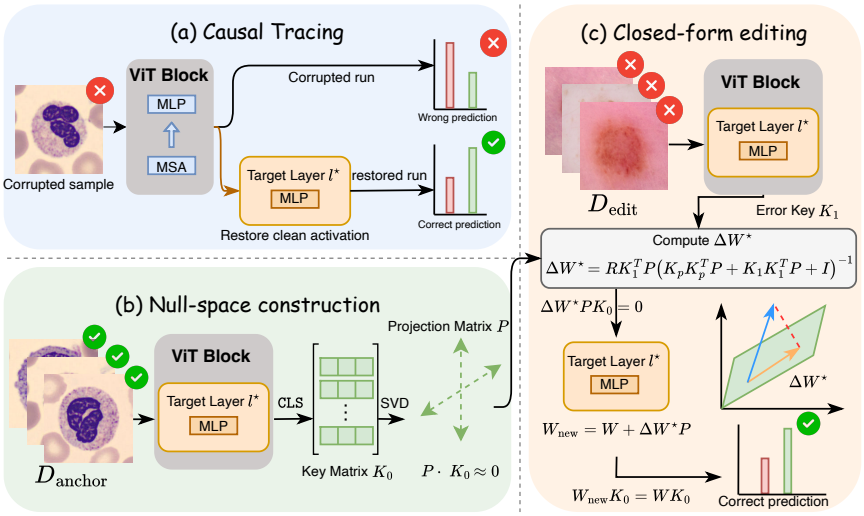

Fig. 2. Overview of X-Edit. We first identify influential layers via causal tracing, then construct a null-space constraint from anchor samples, and finally perform a closed-form parameter update to correct model errors while preserving prior knowledge.

Formally, our objective is to minimize the error on the edit set while strictly preserving the representations of the anchor set as a proxy for test set stability:

$$\begin{aligned} \min_{\Delta\theta} \quad & \mathbb{E}_{(x_e,y_e)\in D_{edit}}[\mathcal{L}(f_{\theta+\Delta\theta}(x_e), y_e)] \\ \text{subject to} \quad & f_{\theta+\Delta\theta}(x_a) = f_\theta(x_a), \quad \forall (x_a, y_a) \in D_{anchor}. \end{aligned} \tag{1}$$

where $\mathcal{L}$ denotes the task-specific loss function.

2.2 Locating Influential Layers via Causal Tracing

Inspired by [14], which demonstrated that feed-forward modules mediate factual associations, we adapt causal tracing to the visual domain to identify the exact location of specific diagnostic knowledge within the ViT.

As shown in Fig. 2(a), given a medical image $x \in D_{edit}$, we first generate a corrupted version $x_{corrupt}$ by injecting Gaussian noise into its embedding to disrupt the prediction. We then perform three forward passes: (1) a **clean run** with x to cache the clean hidden activations $\{h_l^{clean}\}$ across all transformer blocks; (2) a **corrupted run** with $x_{corrupt}$ to obtain the baseline degraded prediction probability; and (3) a **restored run**, where the input remains $x_{corrupt}$, but we artificially restore the MLP activation at a specific layer l back to h_l^{clean} during forward propagation. In the restored run at layer l, we compute the Indirect Effect (IE), defined as the magnitude of recovery in the prediction probability of the correct label. This restoration process is iterated across all transformer

blocks. For each error sample $x \in D_{edit}$, we dynamically identify the top-k layers yielding the highest IE. These identified layers serve as the precise targets for our subsequent null-space parameter editing. For notational simplicity, we collectively denote this sample-specific target set as l^*.

2.3 Null-Space Construction with Anchor Tokens

According to the linear associative memory interpretation of MLPs [4, 8], the weight matrix W of the target layer maps input keys to output values. As shown in Fig. 2(b), to preserve anchor knowledge, we extract the `[CLS]` token input activations to the MLP block at layer l^* for all K samples in D_{anchor}. Stacking these forms the key matrix $K_0 \in \mathbb{R}^{d_{in} \times K}$, where d_{in} is the input dimension.

Projecting the targeted perturbation ΔW into the left null space of K_0 via a projection matrix P satisfies $\Delta W P \cdot K_0 = 0$. Consequently, the updated parameters yield:

$$(W + \Delta W P)K_0 = WK_0 = V_0, \tag{2}$$

where V_0 represents the original output values. This strictly guarantees the preserved knowledge is undisrupted. Computing the null space directly for K_0 is computationally prohibitive for large K. Following AlphaEdit [4], we instead compute it on the non-central covariance matrix $K_0 K_0^\top \in \mathbb{R}^{d_{in} \times d_{in}}$ via SVD:

$$U \Lambda U^\top = \mathrm{SVD}(K_0 K_0^\top). \tag{3}$$

We isolate the null space by removing eigenvectors in U corresponding to non-zero eigenvalues. To ensure numerical stability, we filter out eigenvectors with eigenvalues greater than a negligible threshold, defining the remaining submatrix as $\hat{U}$. The orthogonal projection matrix is thus constructed as $P = \hat{U}\hat{U}^\top$. This geometric constraint completely isolates our editing operations from the anchor representations. Notably, this formulation inherently supports privacy-preserving clinical deployment. Since P relies solely on the pre-computable covariance matrix $K_0 K_0^\top$, the original raw images in D_{anchor} are strictly not required during the online editing phase, circumventing clinical data-availability constraints.

2.4 Closed-Form Update for Sequential Editing

To correct the M samples in D_{edit}, we extract their input activations at layer l^* to form the key matrix $K_1 \in \mathbb{R}^{d_{in} \times M}$. Following the paradigm of [4, 15], we apply several optimization steps to these hidden representations to obtain refined and robust activation targets. Our goal is to find a perturbation yielding target values $V_1 \in \mathbb{R}^{d_{out} \times M}$ that minimize the classification error.

Under the null-space constraint $\Delta W P$, as in Eq. (1), we minimize $\|(W + \Delta W P)K_1 - V_1\|^2$, introducing an L_2 regularization term $\|\Delta W P\|^2$ to prevent perturbation explosion. For sequential clinical deployments, we must also ensure new updates do not overwrite previous edits. Let K_p represent the key matrix

of all previously edited samples. Since past edits are successfully assimilated ($WK_p \approx V_p$), the preservation constraint simplifies to minimizing $\|\Delta W P K_p\|^2$. The comprehensive objective is:

$$\min_{\Delta W} \left(\|(W + \Delta W P)K_1 - V_1\|^2 + \|\Delta W P K_p\|^2 + \|\Delta W P\|^2 \right). \tag{4}$$

Defining the residual error as $R = V_1 - WK_1$ and setting the derivative to zero yields the optimal, closed-form solution:

$$\Delta W^* = R K_1^\top P (K_p K_p^\top P + K_1 K_1^\top P + I)^{-1}. \tag{5}$$

The updated weight matrix is $W_{new} = W + \Delta W^* P$. This analytical solution obviates the need for computationally intensive gradient-based fine-tuning, facilitating post-hoc exceptionally fast interventions. By structurally embedding P and K_p, the framework natively supports scalable sequential editing while mathematically circumventing catastrophic forgetting.

3 Experiments

3.1 Experimental Setup

Baselines. We evaluate our framework against five methods: LWE [18] (state-of-the-art ViT editing), EWC [10] (classical continual learning against forgetting), FineTune, FineTune+L2 Reg, and full Retrain on the entire dataset (an empirical upper bound for editing efficacy and stability).

Datasets. We use five datasets from MedMNIST v2 [17] (Blood, Derma, Organa, and Retina, with Path reserved for parameter analysis; resized to 224×224) and a liver fibrosis benchmark [12] (LiFS with two binary tasks: S1–S3 vs. S4 (sub1), and S1 vs. S2–S4 (sub2)). To prevent data leakage, the edit set exclusively comprises misclassified instances sampled from a disjoint validation split.

Metrics. We assess editing performance across four comprehensive metrics: (1) Test Performance Drop (ΔACC) measures the accuracy degradation on the unseen test set. (2) Edit Fix Ratio represents the proportion of targeted error samples that are successfully corrected. (3) Drop Per Edit (DPE) is a novel metric we propose to quantify the editing trade-off by measuring the collateral test accuracy loss incurred per successfully rectified sample, defined as $\max(0, \Delta\text{ACC}/N_{\text{corrected}})$. (4) Edit Time records the average computational cost required per edit, reflecting the practical efficiency for real-world deployments.

Implementation Details. For causal tracing, we analyze each failure case using 32 corrupted runs by injecting Gaussian noise ($\sigma = 0.1$) into patch tokens. Influential layers are identified by measuring the `[CLS]` token recovery score, with the top-3 layers targeted for editing. For null-space construction, we filter eigenvalues below 10^{-2} and utilize a default anchor set of 500 samples. The target optimization for V_1 consists of 5 iterations with a learning rate of 0.1. All gradient-based baselines are optimized using the Adam optimizer for 10 epochs to ensure convergence on the edit set. For the LWE baseline, a task-specific hypernetwork is pre-trained for each dataset. All experiments are implemented in PyTorch and executed on NVIDIA GeForce RTX 3090 GPUs.

Table 1. Quantitative comparison of editing reliability (measured by test accuracy drop, ΔACC), effectiveness (Fix Ratio), and efficiency (time per sample) across seven medical image datasets. The symbol † denotes baselines that require access to the original pre-training data.

Dataset (N = # Test Cases)	Methods	ΔACC (pp)	Fix Ratio	DPE (Drop/Fix)	Edit Time (Sec/Sample)
bloodmnist (N=3,421)	ReTrain†	↑ 0.29	100.00% (14/14)	0.00	16.44
	FineTune	↓ 0.94	100.00% (14/14)	0.07	1.71
	FineTune+L2	↓ 0.94	100.00% (14/14)	0.07	1.68
	EWC [10]	↓ 1.02	100.00% (14/14)	0.07	1.85
	LWE [18]	↓ 3.27	85.71% (12/14)	0.27	3.03
	X-Edit	↓ 0.36	100.00% (14/14)	0.03	0.53
dermamnist (N=2,005)	ReTrain†	↓ 0.30	100.00% (93/93)	0.00	1.47
	FineTune	↓ 4.04	100.00% (93/93)	0.04	0.35
	FineTune+L2	↓ 4.04	100.00% (93/93)	0.04	0.52
	EWC [10]	↓ 3.84	100.00% (93/93)	0.04	0.56
	LWE [18]	↓ 28.38	82.80% (77/93)	0.37	2.74
	X-Edit	↓ 1.00	98.92% (92/93)	0.01	1.31
organamnist (N=17,778)	ReTrain†	↑ 0.03	100.00% (3/3)	0.00	239.62
	FineTune	↓ 0.03	100.00% (3/3)	0.01	31.83
	FineTune+L2	↓ 0.03	100.00% (3/3)	0.01	24.57
	EWC [10]	↓ 0.06	100.00% (3/3)	0.02	27.96
	LWE [18]	↑ 0.06	100.00% (3/3)	0.00	54.00
	X-Edit	↑ 0.23	100.00% (3/3)	0.00	0.53
retinamnist (N=400)	ReTrain†	↑ 1.50	100.00% (38/38)	0.00	0.63
	FineTune	↓ 3.50	100.00% (38/38)	0.09	0.24
	FineTune+L2	↓ 3.50	100.00% (38/38)	0.09	0.45
	EWC [10]	↓ 4.00	100.00% (38/38)	0.11	0.51
	LWE [18]	↓ 9.25	71.05% (27/38)	0.34	3.62
	X-Edit	↓ 0.50	100.00% (38/38)	0.01	0.91
LiFS-subtask1 (N=141)	ReTrain†	↓ 11.35	90.91% (10/11)	1.14	1.34
	FineTune	↓ 29.79	100.00% (11/11)	2.71	0.45
	FineTune+L2	↓ 29.79	100.00% (11/11)	2.71	0.64
	EWC [10]	↓ 31.21	100.00% (11/11)	2.84	0.74
	LWE [18]	↓ 33.33	54.55% (6/11)	5.56	2.29
	X-Edit	↓ 5.68	100.00% (11/11)	0.52	0.92
LiFS-subtask2 (N=141)	ReTrain†	↓ 1.42	50.00% (4/8)	0.36	1.92
	FineTune	↓ 56.74	100.00% (8/8)	7.09	0.56
	FineTune+L2	↓ 56.74	100.00% (8/8)	7.09	0.73
	EWC [10]	↓ 56.03	100.00% (8/8)	7.00	0.86
	LWE [18]	0.00	12.50% (1/8)	0.00	1.19
	X-Edit	↓ 4.25	100.00% (8/8)	0.53	0.18

3.2 Results

Main results. A clinically viable model editing framework must reliably correct newly encountered errors without degrading the established diagnostic capabilities. As demonstrated in Table. 1, unconstrained gradient-based methods, including FineTune, FineTune+L2, and EWC, suffer from severe catastrophic forgetting. This degradation is particularly devastating in complex, real-world clinical task liver fibrosis staging. For instance, on the LiFS-subtask2 dataset, standard FT leads to a severe performance collapse, with test accuracy dropping by 56.74 percentage points. Similarly, the SOTA ViT editing baseline, LWE, ex-

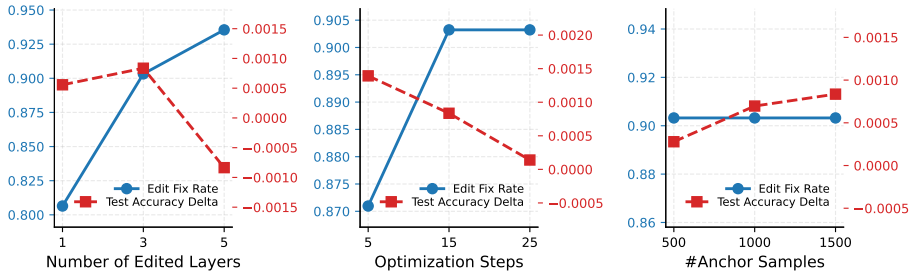

Fig. 3. Parameter analysis for the number of edited Layers, optimization steps and size of anchor sets.

hibits extreme instability, incurring drops of 28.38 and 33.33 percentage points on DermaMNIST and LiFS-subtask1, respectively. In contrast, X-Edit demonstrates exceptional reliability. It incurs near-zero performance drops across all MedMNIST datasets and restricts the degradation on the highly challenging LiFS tasks to merely 5.68 and 4.25 percentage points, respectively, remarkably approaching the empirical upper bound established by ReTrain baseline.

Beyond mere preservation of test stability, the intervention must successfully assimilate the targeted corrections. While standard FT naturally achieves a 100% Fix Ratio by aggressively overwriting parameters, it destroys the base knowledge. Conversely, LWE frequently struggles to learn new corrections on hard cases, yielding poor Fix Ratios (54.55% on LiFS-subtask1 and 12.50% on LiFS-subtask2). X-Edit overcomes this dilemma, consistently achieving near 100% Fix Ratios across almost all benchmarks. Furthermore, the DPE metric explicitly quantifies this trade-off by measuring the collateral damage inflicted on the test set per successfully rectified sample. X-Edit achieves the lowest Drop/Fix ratio among all methods. Furthermore, leveraging its closed-form solution, X-Edit demonstrates high computational efficiency, consistently requiring significantly less editing time than the SOTA ViT editing baseline, LWE.

Parameter Analysis. We evaluate the sensitivity of X-Edit to three key hyperparameters on PathMNIST. As illustrated in Fig. 3, a trade-off exists between the Edit Fix Ratio and Test Performance Drop: (1) Increasing the number of edited layers or optimization steps enhances the fix rate but exacerbates test accuracy degradation. (2) Conversely, expanding the anchor set size reinforces stability by tightening the geometric null-space constraints. These findings empirically validate our theoretical design, highlighting the balance between acquiring new knowledge and preserving established diagnostic representations.

4 Conclusion

In this paper, we proposed **X-Edit**, a null-space model editing framework designed to rectify errors in medical ViTs without catastrophic forgetting. By integrating causal tracing for mechanistic interpretability with a mathematically

transparent, closed-form analytical solution, X-Edit enables exact and efficient post-hoc interventions. Extensive evaluations demonstrate that X-Edit's geometric constraints successfully translate into empirical reliability. By delivering **eXact** updates, **eXplicit** error correction, and an **eXplainable** mechanism, X-Edit provides a step toward controllable post-hoc editing for medical ViTs and can support long-term clinical model adaptation.

Acknowledgments. This work was funded by the Science and Technology Commission of Shanghai Municipality (25TS1412100), the Noncommunicable Chronic Disease-National Science and Technology Major Project (2026ZD0555800/2026ZD0555802), and the National Natural Science Foundation of China (62372115).

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

References

1. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation (2021)
2. Chen, R.J., Wang, J.J., Williamson, D.F., Chen, T.Y., Lipkova, J., Lu, M.Y., Sahai, S., Mahmood, F.: Algorithmic fairness in artificial intelligence for medicine and healthcare. Nature biomedical engineering **7**(6), 719–742 (2023)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021)
4. Fang, J., Jiang, H., Wang, K., Ma, Y., Jie, S., Wang, X., He, X., Chua, T.S.: Alphaedit: Null-space constrained knowledge editing for language models. In: Proceedings of the International Conference on Learning Representations (ICLR) (2025)
5. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Transactions on Biomedical Engineering **69**(3), 1173–1185 (2021)
6. Hacohen, G., Tuytelaars, T.: Predicting the susceptibility of examples to catastrophic forgetting. In: Proceedings of the International Conference on Machine Learning (ICML) (2025)
7. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: MICCAI brainlesion workshop. pp. 272–284 (2021)
8. Hu, C., Cao, P., Chen, Y., Liu, K., Zhao, J.: Wilke: Wise-layer knowledge editor for lifelong knowledge editing. In: Findings of the Association for Computational Linguistics (ACL). pp. 3476–3503 (2024)
9. Huang, X., Zhao, K., Huang, L.K.: Model editing for vision transformers. In: Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS) (2025)
10. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)

11. Lekadir, K., Frangi, A.F., Porras, A.R., Glocker, B., Cintas, C., Langlotz, C.P., Weicken, E., Asselbergs, F.W., Prior, F., Collins, G.S., et al.: Future-ai: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. bmj **388** (2025)
12. Liu, Y., Gao, Z., Shi, N., Wu, F., Shi, Y., Chen, Q., Zhuang, X.: Merit: Multi-view evidential learning for reliable and interpretable liver fibrosis staging. Medical Image Analysis **102**, 103507 (2025)
13. Liu, Y., Zhen, R., Gao, S., Luo, X., Gao, X., Chen, Q., Zhuang, X.: Bayesmm: Robust deep combined computing tackling heavy-tailed distribution in medical images. In: Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 45–54 (2025)
14. Meng, K., Bau, D., Andonian, A., Belinkov, Y.: Locating and editing factual associations in gpt. In: Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS) (2022)
15. Meng, K., Sharma, A.S., Andonian, A., Belinkov, Y., Bau, D.: Mass-editing memory in a transformer. In: Proceedings of the International Conference on Learning Representations (ICLR) (2023)
16. Van der Velden, B.H., Kuijf, H.J., Gilhuijs, K.G., Viergever, M.A.: Explainable artificial intelligence (xai) in deep learning-based medical image analysis. Medical image analysis **79**, 102470 (2022)
17. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data **10**(1), 41 (2023)
18. Yang, Y., Huang, L.K., Chen, S., Ma, K., Wei, Y.: Learning where to edit vision transformers. In: Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS). vol. 37, pp. 134459–134487 (2024)
19. Zeng, G., Chen, Y., Cui, B., Yu, S.: Continual learning of context-dependent processing in neural networks. Nature Machine Intelligence **1**(8), 364–372 (2019)