Fusing Parametric and Spatial Constraints in Continual Learning

Giacomo Ignesti
Institute of Science and Information Technologies
National Research Council
Pisa, Italy

Gennaro D'Angelo
Institute of Clinical Physiology
National Research Council
Pisa, Italy

Lorenza Pratali
Institute of Clinical Physiology
National Research Council
Pisa, Italy

Massimo Martinelli
Institute of Science and Information Technologies
National Research Council
Pisa, Italy

***Abstract*—The deployment of powerful foundation models in medical imaging is severely hampered by the fact that neural networks cannot learn sequentially from decentralised data without catastrophic forgetting. Early continual learning frameworks, such as Elastic Weight Consolidation (EWC), offer a parametric defence against forgetting but treat the network as a black box and neglect to preserve learned spatial representations. To solve this problem, we propose a novel continual learning framework that anchors EWC within a stable reconstruction-classification training paradigm using batch-trained prototypes. Our approach fundamentally constrains representation drift by enforcing spatial alignment between the network's dynamically generated saliency maps and static visual class prototypes. Our approach combines weight consolidation with spatial reconstruction penalties to explicitly prevent the prior classification state-space from being distorted during novel task updates. Empirical evaluations show that the fusion not only retains high classification efficacy on historical tasks but also successfully maintains the network's visual interpretability. Ultimately, this framework establishes a new direction for continual learning in medical imaging, shifting the focus from purely weight-based regularisation to the holistic preservation of both classifier stability and interpretable feature representations.**

***Index Terms*—AI, CL, XAI, Imaging, Ultrasounds, Learning**

I. Introduction

Classical mathematical and physical models evolve; the governing formulas generally change to preserve prior knowledge as a limiting case. A quintessential example of this is the transition from classical mechanics to quantum mechanics. While the underlying postulates of quantum theory represent a radical paradigm shift, classical mechanics is not discarded; rather, it emerges seamlessly as a macroscopic limit [1]. A similar cumulative principle applies to classical statistical and probabilistic algorithms. In these frameworks, particularly within Bayesian inference, new data and observations continuously refine the model's understanding without erasing the past. While historical data may become statistically less relevant to current trends over time, it is nevertheless mathematically integrated into the posterior distribution without compromising the model's fundamental mechanics. However, these robust, cumulative properties do not faithfully translate to modern Artificial Intelligence (AI) models based on Deep Learning (DL). The highly non-convex nature of neural network optimisation and the lack of closed-form constructive solutions mean that a fully trained model cannot simply assimilate new information sequentially. This limitation manifests as *catastrophic forgetting*, a phenomenon wherein retraining a DL model on a new, continuous data stream causes an abrupt and inherent loss of previously acquired knowledge [2]. This challenge is fundamentally rooted in the *stability-plasticity dilemma*: the intrinsic tension between maintaining stability (retaining old knowledge) and exhibiting plasticity (learning new patterns) [7]. Mathematically, this occurs because the gradient descent steps taken during retraining move the model's weight matrices in a direction optimised solely for the new data. Unless strictly constrained, such as by projecting weight updates to be orthogonal to the feature spaces of previous tasks, these gradient steps forcefully push the weights out of the optimal basins for past tasks, overwriting historical memory rather than integrating it [6].

This issue could theoretically be avoided by retraining the entire model on an expanded, updated dataset. However, this approach imposes significant constraints on computational power, time, and data availability. Furthermore, there is no guarantee that the retrained model will maintain its previous performance, which is why the aforementioned alternative solution is preferred. It is important to highlight that in the medical imaging domain, data availability is a critical factor. Given strict privacy regulations and the limited usability of historical data, enabling this new training modality is particularly important, as its benefits are clear [12].

Consequently, to validate algorithms that can learn sequentially without relying on exhaustive historical data, the research community relies on standardised Continual Learning (CL) benchmarks. These benchmarks evaluate models across distinct, incremental tasks, measuring not just plasticity on

novel data, but strictly quantifying memory retention through a backwards-transfer evaluation matrix ($R_{i,j}$).

This paper proposes a novel solution to catastrophic forgetting, as well as the broader issues of model interpretability and representational drift, by formalising a dual-penalty training modality. We introduce a methodology that enforces both parameter-space stability and function-space alignment across sequential tasks. To achieve parameter-space stability, we utilise Elastic Weight Consolidation (EWC) to compute the Fisher Information Matrix for each learned task. This isolates and protects the specific weights critical to past distributions, ensuring that subsequent gradient updates do not destabilise the network's foundational parameters. Because protecting weights alone is insufficient when the model's visual logic shifts, we also achieve function-space alignment through Batch-CAM prototypes [5]. To prevent representational drift, we construct spatial prototypes of previously learned classes. During the acquisition of new tasks, a reconstruction loss is paired with a categorical loss to enforce spatial alignment between the new Class Activation Maps (CAM) and the historical prototypes.

This approach forces the model to retain its technical visual focus on meaningful features while preventing the manifold of the encoded class distribution from deviating. However, enforcing such strict spatial and parametric integrity introduces a secondary challenge, particularly acute in medical imaging, where sequential classes often represent a continuous pathological spectrum (e.g., the escalation of vertical B-line coalescence in lung ultrasounds).

Under robust dual-penalty regularisation, the fully connected classification head can succumb to task bias or catastrophic intransigence—failing to map newly acquired but structurally overlapping features to their correct output neurons, thereby protecting historical pathways. To rigorously isolate and address this bottleneck, our methodology employs a progressive manifold stability-tracking framework.

By utilising Structural Similarity Index Measure (SSIM) tracking on dynamic activation maps, alongside continuous Linear Probing, we aim to mathematically decouple the performance of the convolutional feature extractor from that of the linear classification head. Consequently, to fully exploit the pristine and highly separable latent representations preserved by our dual-penalty backbone, we propose bypassing the standard fully connected layer entirely.

The subsequent sections of this paper are structured to empirically validate this hypothesis and methodology. Section II introduces two inspired standardised Class-Incremental benchmarking datasets, Split-FashionMNIST [4] and Split-POCUS, detailing the pre-processing protocols required for dynamic visual mapping along a pathological spectrum. Section III introduces the DL architecture employed, while Section IV outlines the proposed dual-penalty architecture, detailing the mathematical interplay among EWC, prototype-alignment losses, and manifold stability metrics. Finally, Section V presents our experimental results, followed by Sections VI-VII, which highlight the criticality and advances of the proposed method, as well as new directions for future work in continuous medical diagnostics.

II. Data

The subsequent experiments evaluating CL adaptations were conducted using the widely recognised Fashion-MNIST (F-MNIST) dataset alongside a proprietary lung ultrasound (LUS) image dataset. To establish a robust Class-Incremental Learning baseline, Fashion-MNIST is structured into the standard "Split-FMNIST" benchmark. In this setup, the 10 categorical classes are divided into five sequential tasks, each consisting of two classes. Following validation on the Split-FMNIST benchmark, the exact same incremental CL methodology is applied to the clinical "Sonographer task" using a proprietary LUS dataset. This dataset comprises lung sonograms acquired from 30 unique subjects. To ensure consistency and eliminate hardware-induced variability, all acquisitions were performed using the same ultrasound machine, clinical methodology, and acquisition protocol [15].

The temporal distribution of the data collection provides a natural division for CL scenarios. Data from 25 subjects were collected during the initial phase, while data from the remaining 5 were recorded two years later. In line with the incremental learning perspective established in the F-MNIST experiments, we denote the initial data cohort as **Task 1** and the subsequent cohort as **Task 2**. From the raw ultrasound videos of these 30 subjects, a total of 2,000 distinct frames were extracted. These frames were strictly annotated by medical professionals into four distinct clinical classes representing a gradient of lung pathology. The severity of the condition is quantitatively assessed by the presence and density of B-lines in the pleural space, ranging from a healthy lung profile (no B-lines) to severe impairment. To ensure reproducibility and standardisation, the class criteria are defined by the percentage of the image area covered by B-lines, as detailed in Table I.

TABLE I
Distribution of Lung Pathologies and B-Line Coverage

Class	Condition Label	B-Line Area Coverage
1	Healthy Lung	0%
2	Mild Impairment	$< 30\%$
3	Moderate Impairment	$30\% - 70\%$
4	Severe Impairment	$> 70\%$

Rigorous data pre-processing and augmentation pipelines were established to enhance model robustness while preserving critical features across both datasets. For the LUS images, data augmentation was strictly limited to basic linear geometric transformations—specifically, slight rotations and minimal shear variations—that simulate natural probe orientation shifts without distorting the underlying medical morphology. Photometric transformations, such as changes in pixel brightness or contrast, were deliberately omitted to prevent the introduction of artefacts that the model might misinterpret as pathological features. To maintain methodological consistency, a related,

strictly geometric pipeline is applied to the F-MNIST dataset. Following augmentation, all images from both datasets were tensorized and scaled to a $[0, 1]$ interval. Finally, grayscale channel normalisation was applied. Crucially, the normalisation statistics (mean and standard deviation) were computed independently for each sequential task (e.g., Task 1 and Task 2 for LUS, and Tasks 1 through 5 for F-MNIST). This isolated normalisation strategy is vital in a CL setup, given the inherent data spread and shifting distributions across acquisition periods, thereby accurately reflecting the real-world domain shifts that occur over time.

III. Model Selection

All experimental evaluations were conducted utilising the ConvNeXt-Tiny [9] architecture, instantiated with the standard PyTorch libraries. To tailor the model for these specific imaging tasks, critical structural modifications were introduced to the input and output stages. The initial processing pipeline was adapted to handle single-channel grayscale sonograms and standard dataset inputs by mapping them to the network's required dimensionality while preserving spatial integrity. Furthermore, the fully connected (FC) classification head was heavily truncated relative to the standard ConvNeXt configuration to directly correspond to the distinct classes in each experiment. This resulted in an output layer of 10 neurons for the Fashion-MNIST baseline tasks and 4 neurons representing the clinical severity classes in the lung ultrasound task.

Beyond the structural adaptations, the network was optimised utilising a dual-penalty Continual Learning paradigm. This strategy preserves the core block structure and weights of the ConvNeXt architecture by intervening primarily through specialised regularisation techniques. To ensure parameter-space stability, two techniques are employed: EWC for FC layer parameter stability and, concurrently, function-space alignment via a prototypical reconstruction learning strategy. This augmented loss employs a contrastive formulation that explicitly aligns the network's spatial feature representations with predefined class-specific prototypes.

To achieve this alignment, the gradient-weighted Class Activation Maps (Grad-CAM) [14] of the target images are dynamically generated during training to force the network to localise and focus on technically meaningful details. Crucially, the contrastive comparison in the loss formulation is measured using SSIM and not computed on a per-instance basis. Instead, it is aggregated and evaluated at the batch level for each respective class distribution, building upon the methodological framework introduced in our prior work. This batch-by-batch distribution mapping ensures that the model learns a stable and interpretable representation of the data manifold without deviating from the previously introduced Batch-CAM concept.

IV. Continual Learning Methodology and Hybrid Regularisation

The baseline for our continual learning experiments is the EWC methodology, which addresses the stability-plasticity dilemma and mitigates catastrophic forgetting as outlined in Section I.

The core mechanism of EWC relies on the Fisher Information Matrix to establish a mathematical connection between the network states of previously learned tasks and those of the subsequent task. Specifically, the diagonal of the Fisher Information Matrix, denoted as F_i, is utilised to quantify the importance of each specific parameter for the previous predictions. When updating the network for the new dataset, this matrix enables the model to maintain stability by penalising deviations in weights that were crucial to the historical data. The standard continual learning loss constraint under Elastic Weight Consolidation is formalised as:

$$\mathcal{L}_{CL}(\theta) = \mathcal{L}_{\mathcal{D}_B}(\theta) + \sum_i \frac{\lambda_{EWC}}{2} F_i(\theta_i - \theta^*_{A,i})^2 \qquad (1)$$

where $\mathcal{L}_{\mathcal{D}_B}(\theta)$ represents the standard Cross Entropy loss on the newly acquired dataset, $\theta^*_{A,i}$ denotes the optimal weights derived from the previous task, and λ_{EWC} serves as the regularization coefficient controlling the mathematical trade off between plasticity and stability.

While EWC provides a robust foundation for parameter stability, it protects only the model's weights and does not inherently preserve spatial reasoning or explainability. To address this limitation, we augment the traditional parameter update strategy with a novel Prototype Rehearsal approach based on the Batch-CAM methodology introduced.

In this proposed hybrid framework, we couple the weight-consolidation term with an additional regularisation term that focuses on spatial attention. Rather than utilising traditional replay strategies that require retaining raw historical images, which introduce significant privacy and storage concerns, the system strictly retains Class Prototypes denoted as P_c. These prototypes represent the ideal activation maps for specific visual features, such as canonical vertical artefact patterns in sonograms or topological structures in standard dataset mapping.

During retraining on a new sequential task, these stored prototypes serve as active spatial regularisers. By aligning the network's current activation maps with the historical prototypes, we prevent the model's internal attention mechanisms from drifting. Consequently, this dual penalty approach ensures that the sequentially trained model not only maintains high classification accuracy across all learned tasks but also preserves a consistent and interpretable reasoning process. This methodology effectively anchors the network's focus [16] on technically meaningful visual details throughout the model's lifespan.

The spatial prototype regularisation loss is defined as:

$$\mathcal{L}_P(\theta) = \frac{1}{|\mathcal{C}_B|} \sum_{c \in \mathcal{C}_B} (1 - \text{SSIM}(M_c(x;\theta), P_c)) \qquad (2)$$

where $\mathcal{C}_B$ represents the set of unique classes present in the current training batch, $M_c(x;\theta)$ denotes the dynamically

generated Batch CAM for class c given the current network parameters θ and input x, and P_c is the corresponding frozen historical prototype.

To present the complete hybrid methodology, the finalised total loss function combines the categorical classification, the parameter space weight consolidation, and the function space spatial regularisation terms:

$$\mathcal{L}_T(\theta) = \mathcal{L}_{\mathcal{D}_B}(\theta) + \frac{\lambda_{EWC}}{2} \sum_i F_i(\theta_i - \theta^*_{A,i})^2 + \lambda_P \mathcal{L}_P(\theta) \quad (3)$$

where λ_P serves as the hyperparameter governing the strength of the function space alignment in the prototype space.

Enforcing strict spatial and parametric integrity introduces a secondary challenge, particularly acute in continuous pathological spectrums (e.g., the progressive coalescence of B-lines in sonograms). Under robust dual-penalty regularisation, the FC classification head can succumb to task bias or *catastrophic intransigence*—mathematically refusing to map structurally overlapping visual features to new output neurons in order to protect historical pathways.

To rigorously isolate this classifier bottleneck, our methodology incorporates a Progressive Manifold Stability tracking framework. We mathematically decouple the performance of the convolutional feature extractor, denoted as $\phi(x; \theta_{conv})$, from the linear classification head, utilising three specific reference metrics: Latent Centroid Cosine Similarity, Mean Structural Similarity Index Measure (MSSIM) distance, and Linear Probing.

To quantify feature-space plasticity and geometric cluster shifting independently of standard classification accuracy, we evaluate the representation in the latent space immediately preceding the linear layer. The latent centroid $\mu_c^{(t)}$ for class c after learning task t is defined as the mean feature representation:

$$\mu_c^{(t)} = \frac{1}{|\mathcal{D}_c|} \sum_{x \in \mathcal{D}_c} \phi(x; \theta_{conv}^{(t)}) \quad (4)$$

The geometric stability of the encoded manifold is continuously monitored using Cosine Similarity between the baseline centroid (at the time the task t_0 was initially learned) and the current centroid:

$$\mathcal{S}_c^{(t)} = \frac{\mu_c^{(t_0)} \cdot \mu_c^{(t)}}{\|\mu_c^{(t_0)}\| \|\mu_c^{(t)}\|} \quad (5)$$

This metric explicitly identifies whether the internal data manifold is warping sequentially or if feature clusters remain geometrically anchored.

While SSIM is utilised as a training loss, we also employ the Mean Structural Similarity Index Measure (MSSIM) as an independent evaluation distance metric to quantify *attention drift*. By anchoring a baseline instance-level CAM immediately after a task is learned, we measure the structural degradation of the model's visual logic after subsequent tasks. A sustained high MSSIM score proves that the convolutional filters maintain their focus on technically meaningful geometries (e.g., specific artefacts) rather than devolving into random background activations.

To definitively prove that the feature representations remain perfectly separable despite potential catastrophic drops in standard continual learning accuracy, we deploy a Linear Probing evaluation protocol. After the continuous learning pipeline concludes, the entire convolutional backbone θ_{conv} is strictly frozen. An independently initialised linear classification head is then trained on the preserved feature space. Near-perfect recovery of classification accuracy with this linear probe mathematically guarantees that catastrophic forgetting within the model is entirely localised to the linear layer's task bias, thereby validating that the underlying feature extractor successfully captured and retained the distributions across all sequential tasks.

A. Ablation Study Design

To scientifically isolate the efficacy of the proposed spatial Batch CAM alignment, an ablation study was conducted. In this configuration, the spatial penalty is entirely removed by setting $\lambda_P = 0$ during the sequential training phases. The model is consequently optimised, relying strictly on the standard EWC parameter constraint, Eq.1.

During this ablation evaluation, the structural similarity (MSSIM distance) between instance-level activation maps is continuously tracked relative to their pre-shift baseline. A precipitous decline in MSSIM under this protocol provides definitive mathematical proof that, in the absence of $\mathcal{L}_{Proto}$, parameter consolidation alone is fundamentally incapable of preventing spatial attention drift and function-space degradation in neural networks.

V. Results

To balance the stability-plasticity dilemma within the dual-penalty framework, the regularisation hyperparameters were established at $\lambda_{EWC} = 2700$ for parameter space consolidation, $\lambda_P = 1500$ for categorical prototype rehearsal, and $\lambda_{CAM} = 5.0$ to govern the spatial MSSIM alignment constraint.

The parameters were empirically fit using different pairs between 0 and 10000 for CL stability, while for prototype resemblance, values between 0 and 10 were used. The reported configuration was chosen to guarantee the optimal trade-off between retaining historical knowledge (stability) and successfully assimilating new distributions (plasticity) throughout the sequential tasks.

Unconstrained baseline models demonstrated expected catastrophic forgetting, with retention of prior knowledge degrading substantially upon the introduction of new tasks. While the isolated application of EWC successfully restored parameter stability, tracking metrics revealed it failed to maintain spatial alignment without additional constraints.

In the 5-task Split-FashionMNIST benchmark, standard sequential evaluation of the proposed dual-penalty model yielded

lower final classification accuracies on historical tasks, culminating in 48.85% for Task 1 and 22.55% for Task 2. However, Linear Probing on the frozen convolutional backbone resulted in a near-perfect recovery of feature separability, achieving final probe accuracies of 94.30%, 91.75%, 97.85%, 99.30%, and 96.75% across Tasks 1 through 5, respectively. In terms of state space alignment, the model maintained a stable structural similarity index (CAM-SSIM), tracking between 0.57 and 0.74 for Task 1 throughout the continuous training lifecycle, while Latent Centroid Cosine Similarity scores remained anchored above 0.35. In the corresponding ablation study, where the spatial prototype penalty was entirely removed ($\lambda_P = 0$), Linear Probing accuracy [10] remained high ($\geq 98.35\%$), but the Task 1 CAM-SSIM rapidly degraded to 0.4401, and the Latent Centroid Cosine Similarity shifted severely toward orthogonality (0.1207).

In the clinical lung ultrasound evaluation, restricted strictly to the continuous pathological spectrum (Task 1: Mild, Task 2: Moderate, Task 3: Severe), the standard continual learning classification head exhibited severe rigidification. At the conclusion of the sequential training pipeline, standard evaluation yielded 50.34% accuracy on Task 1, 97.37% on Task 2, and 0.00% on Task 3. Conversely, the Linear Probing evaluation achieved exactly 100.00% feature separability and classification accuracy across all three pathological severity tasks simultaneously. Function space alignment was rigorously preserved throughout this sequence, with the Task 1 CAM-SSIM recorded at 0.5760 upon the completion of Task 3.

VI. Conclusion

In medical diagnostics, the imperative of an AI system extends far beyond raw statistical accuracy; it must maintain rigorous spatial interpretability across sequential updates. This study was initially motivated by foundational continual learning experiments utilising simple image reconstruction baselines. These early tests revealed a critical vulnerability: while standard parameter-space constraints mitigate catastrophic forgetting numerically, they fail to prevent severe spatial attention drift.

The prototypical Batch-CAM approach instead seems to focus the network on the right details in successive CL training datasets, as shown in Figure 1.

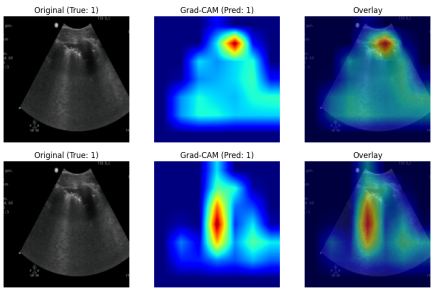

Fig. 1. Comet-tail highlighted in the naïve CL approach in which the test set of task 1 seems to be analysed more correctly after task 2

This initial observation prompted the subsequent transition to real-world clinical benchmarks, specifically mapping the continuous pathological spectrum of lung ultrasound artefacts. The implications of these advanced experiments are profound. Qualitative monitoring via activation maps, alongside continuous quantitative tracking (as summarised in Table II), confirmed that unconstrained or singly-constrained models rapidly lose focus on critical diagnostic regions, resorting to uninterpretable background geometries. Conversely, the proposed hybrid dual-penalty methodology consistently anchored the model's technical focus on clinically relevant features, such as pleural lines and vertical B-line coalescences. These findings establish that integrating prototypical spatial rehearsal is not merely an augmentation but an absolute necessity for enforcing interpretability and ensuring that a model's diagnostic logic remains resilient over its lifespan, Figure 2.

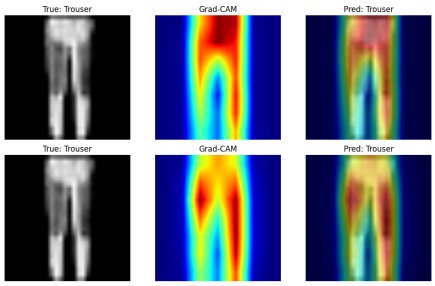

Fig. 2. An F-MINST example of task 1 prototype extraction after the model is updated on task 5, showing remarkable detail extraction capacity even without model accuracy on the same task.

TABLE II
THE TABLE CONTRASTS STANDARD CLASSIFIER FAILURE AGAINST THE PRISTINE FEATURE SEPARABILITY PRESERVED IN THE BACKBONE, VALIDATED BY SUSTAINED CAM-SSIM.

Evaluation Metric	Task 1 (Mild)	Task 2 (Moderate)	Task 3 (Severe)
Standard CL Accuracy	50.34%	97.37%	0.00%
Linear Probing Accuracy	100.00%	100.00%	100.00%
Spatial Alignment (CAM-SSIM)	0.5760	0.3789	Baseline

VII. Future Directions

To further advance continuous model evolution, future architectures must fundamentally decouple the mechanisms of catastrophic forgetting. Historically, severe task-recency bias has been attributed primarily to the FC classification layer. During sequential learning, this linear layer suffers from an extreme score-magnitude bias, overwhelmingly favouring newly introduced classes and effectively suppressing historical outputs [8]. Consequently, modern architectural interventions often seek to bypass or heavily penalise the classifier head entirely, occasionally replacing static FC layers with dynamically growing self-organising networks to prevent the aggressive forgetting typical of standard classifiers [11].

However, isolating the fault solely in the linear classifier overlooks representational drift in the underlying convolutional feature extractor. Bias and structural degradation are

fundamentally shared across the entire architecture during class-incremental learning [17]. As convolutional filters update to accommodate novel tasks, the optimal mathematical space utilised to represent prior distributions becomes severely warped. This structural overwriting destroys the historical principal directions necessary to separate older classes, ultimately eliminating linear separability within the latent space [3].

Therefore, future continuous learning frameworks must deploy unified dual strategies: explicitly preventing feature-extractor drift while simultaneously neutralising linear-classifier bias [8]. Building upon the empirical success of our Linear Probing evaluations, future iterations of this work will transition the proposed hybrid-regularised backbone into a fully non-parametric evaluation paradigm. By discarding the biased FC layer in favour of a dynamic Nearest Class Mean (NCM) [13] mapping system based on protected latent prototypes, the framework can scale organically. Furthermore, stabilising the backbone while removing the FC bottleneck provides an optimal foundation for decentralised federated protocols, enabling edge devices to continuously compute and transmit gradients for new clinical distributions without compromising historical stability or patient privacy.

REFERENCES

[1] Paul Adrien Maurice Dirac. *The Principles of Quantum Mechanics*. Clarendon Press, Oxford, UK, 4th revised edition, 1981.

[2] Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999.

[3] Qiao Gu, Dongsub Shim, and Florian Shkurti. Preserving linear separability in continual learning by backward feature projection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24286–24295, 2023.

[4] Yen-Chang Hsu, Yen-Cheng Liu, Anita Ramasamy, and Zsolt Kira. Re-evaluating continual learning scenarios: A categorization and case for strong baselines. *arXiv preprint arXiv:1810.12488*, 2018.

[5] Giacomo Ignesti, Davide Moroni, and Massimo Martinelli. Batch-cam: Introduction to better reasoning in convolutional deep learning models. *arXiv preprint arXiv:2510.00664*, 2025.

[6] 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.

[7] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

[8] Xialei Liu, Chenshen Wu, Mikel Menta, Luis Herranz, Bogdan Raducanu, Andrew D Bagdanov, Shangling Jui, and Joost van de Weijer. Generative feature replay for class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 226–227, 2020.

[9] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[10] Zheda Mai, Ruiwen Li, Hyunwoo Kim, and Scott Sanner. Supervised contrastive replay: Revisiting the nearest class mean classifier in online class-incremental continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3589–3599, 2021.

[11] Duvindu Piyasena, Miyuru Thathsara, Sathursan Kanagarajah, Siew Kei Lam, and Meiqing Wu. Dynamically growing neural network architecture for lifelong deep learning on the edge. In *2020 30th International Conference on Field-Programmable Logic and Applications (FPL)*, pages 262–268. IEEE, 2020.

[12] Mohammad Areeb Qazi, Anees Ur Rehman Hashmi, Santosh Sanjeev, Ibrahim Almakky, Numan Saeed, Camila Gonzalez, and Mohammad Yaqub. Continual learning in medical imaging: A survey and practical analysis. *ACM Computing Surveys*, 58(8):1–25, 2026.

[13] 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*, pages 2001–2010, 2017.

[14] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[15] Giovanni Volpicelli, Luna Gargani, Stefano Perlini, Stefano Spinelli, Greta Barbieri, Antonella Lanotte, Gonzalo García Casasola, Ramon Nogué-Bou, Alessandro Lamorte, Eustachio Agricola, et al. Lung ultrasound for the early diagnosis of covid-19 pneumonia: an international multicenter study. *Intensive care medicine*, 47(4):444–454, 2021.

[16] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[17] Fei Zhu, Xu-Yao Zhang, and Cheng-Lin Liu. Calibration for non-exemplar based class-incremental learning. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.