UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Dottorato di ricerca in
"Information and Communication Technologies (ICT)"

Ciclo XXXVIII

Evolving Knowledge in Artificial Intelligence: Toward Robust and Modular Deep Neural Networks

Giacomo Capitani

Relatore: Prof. Elisa Ficarra
Correlatore: Prof. Simone Calderara

Coordinatore del Corso di Dottorato: Prof. Luigi Rovati

Review committee:

Prof. Concetto Spampinato, University of Catania
Prof. Emanuele Rodolà, Sapienza University of Rome

To my family and friends

Abstract

Deep neural networks have become a cornerstone of modern Artificial Intelligence due to their remarkable effectiveness and versatility. However, their performance can deteriorate across diverse learning scenarios: models may rely on shortcut solutions in the presence of spurious correlations, suffer from catastrophic forgetting when data streams evolve over time, and exhibit limited compositional or knowledge transfer abilities. The present thesis explores how neural models can be directed to adapt, preserve, transfer, and compose their capabilities in such non-stationary scenarios by leveraging techniques that go beyond naïve data fitting, which would fail to remain effective under these conditions.

The first part focuses on mitigating shortcut learning. We show how, in the absence of explicit protected attributes, latent clusters can be leveraged to form proxy groups that guide optimization trajectories away from malign solutions, thereby improving robustness. The analysis is then extended to continual learning, where rehearsal-based strategies have been discovered to introduce or amplify spurious correlations. To address this issue, a rehearsal mechanism is proposed to maintain a balanced buffer with respect to loss values and mitigate forgetting under distributional shifts. Additionally, the thesis investigates multimodal vision–language models and shows that CLIP-like architectures exhibit implicit, human-analogous biases, which are characterized by adapting established measurement tools from social psychology.

The second part investigates how the parameter space of neural networks can be manipulated to enhance post-training knowledge transfer. First, it examines when task vectors retain transferable knowledge across models trained on distinct datasets and introduces a permutation-based alignment algorithm for transformer-based models. Finally, the geometric properties of the loss landscape, particularly its flatness, are shown to predict compatibility when merging multiple fine-tuned models derived from a common pre-training, with practical applications in 3D medical image segmentation. Extensive experimental analyses across diverse datasets and learning paradigms support these findings.

Collectively, the contributions outline a three-axis framework: (i) mitigation of shortcut learning; (ii) avoidance of spurious correlations in continual learning scenarios; (iii) manipulation of parameter-space geometry for knowledge transfer and model merging. Through this perspective, the thesis advances principles and methodologies for developing AI systems whose capabilities can be robustly maintained, transferred, and composed.

Contents

Book I

Introduction

1. Notation

This chapter provides a comprehensive summary of the mathematical notation, symbols, and conventions adopted throughout this thesis.

General Mathematical Symbols

- $\mathbb{R}$: set of real numbers.
- $\mathbb{N}$: set of natural numbers.
- $p(x)$: probability distribution over random variable x.
- $p(y|x)$: conditional probability distribution.
- $\mathbb{E}_{\mathrm{x}\sim\mathrm{p(x)}}[f(x)]$: expected value of function $f(x)$ with respect to distribution $p(x)$.
- $\mathbb{I}[\cdot]$: indicator function, equals 1 if the condition is true, 0 otherwise.

Linear Algebra

- $\mathbf{A}$, $\mathbf{B}$: matrices in $\mathbb{R}^{m\times n}$.
- $\mathbf{v}$, $\mathbf{w}$: vectors in $\mathbb{R}^n$.
- $\mathbf{A}^\mathsf{T}$: transpose of matrix $\mathbf{A}$.
- $\mathbf{A}^{-1}$: inverse of matrix $\mathbf{A}$.
- $\mathbf{A} \odot \mathbf{B}$: element-wise product.
- $\mathbf{A}^\mathsf{T}\mathbf{B}$: matrix multiplication.
- $\langle \mathbf{v}, \mathbf{w} \rangle$ or $\mathbf{v}^\mathsf{T}\mathbf{w}$: inner product between vectors.
- $||\mathbf{v}||_p$: p-norm of vector $\mathbf{v}$. When p is omitted, it denotes the Euclidean norm ($p = 2$).
- $\mathbf{I}$ or $\mathbf{I}_n$: identity matrix of size $n \times n$.

Activation Functions and Operations

- $\sigma(x) = \frac{1}{1+e^{-x}}$: sigmoid (logistic) function.
- $\text{ReLU}(x) = \max(0, x)$: rectified linear unit.
- $\text{softmax}(\mathbf{z})_i = \frac{e^{z_i}}{\sum_j e^{z_j}}$: softmax function applied to vector $\mathbf{z}$.
- $\text{argmin}_x f(x)$: argument x that minimizes function $f(x)$.
- $\text{argmax}_x f(x)$: argument x that maximizes function $f(x)$.
- $(f \circ g)(x)$ or $f(g(x))$: composition of functions f and g.

2. Overview and Motivation

Deep Neural Networks (DNNs) have become the foundation of modern artificial intelligence, achieving remarkable success across diverse domains ranging from computer vision to natural language processing [45, 183, 69, 174, 43, 144]. However, their performance often deteriorates under realistic deployment conditions: models may rely on spurious correlations and shortcut solutions when confronted with distribution shifts [63], suffer from catastrophic forgetting when data streams evolve over time [121], and exhibit limited ability to transfer or compose knowledge across tasks and architectures [82]. This thesis addresses these fundamental limitations through an integrated perspective that examines how DNNs can be made more robust, adaptive, and modular.

Shortcut Learning and Robust Generalization. Deep neural networks frequently exploit *spurious correlations* [63]: patterns that are predictive within the training distribution but fail to capture the underlying causal structure of the task. Such shortcuts may arise from background textures, imaging artifacts, or demographic attributes, and disproportionately harm minority or underrepresented subpopulations [17, 40, 165]. Traditional debiasing approaches assume access to protected group annotations such as gender or ethnicity labels, which are often unavailable or ethically problematic due to privacy concerns and legal constraints [97]. Our foundational hypothesis is that self-supervised feature representations encode latent structure that partially aligns with vulnerable subgroups. In **Chapter 4**, we introduce *ClusterFix*, a two-stage framework that clusters self-supervised embeddings independently of the downstream task and leverages cluster assignment errors as proxies for bias-prone samples. By reweighting these instances during training, we achieve worst-group accuracy improvements that match or exceed those of methods that require protected group labels.

We extend our investigation to *multimodal models*. In **Chapter 6**, we demonstrate that CLIP-like architectures [144] exhibit implicit biases analogous to those measured in human cognition through psychological tools such as the Stereotype Content Model [52] and Implicit Association Tests [136].

Catastrophic Forgetting and Robustness Over Time. While most debiasing interventions are evaluated in offline settings, realistic deployments require models to adapt continuously as new data, domains, or tasks emerge. Under such *continual learning* regimes, neural networks exhibit *catastrophic forgetting* [121]: performance on previously learned tasks degrades sharply as new knowledge is acquired. This temporal instability raises a critical question: can robustness to spurious correlations be maintained as models and data evolve sequentially?

In **Chapter 5**, we demonstrate that standard rehearsal-based continual learning methods, which rely on small memory buffers storing past examples, fail to preserve debiasing properties. Random sampling strategies for buffer population tend to overrepresent majority patterns and amplify spurious correlations, effectively poisoning the replay mechanism. To counteract this degradation, we introduce *Learning without Shortcuts* (LwS), a framework that encourages anti-shortcut representations during current task training, and a loss-based buffer sampling strategy that ensures balanced representation of bias-aligned and bias-conflicting examples. This dual approach enables continual learning systems to maintain worst-group accuracy across task sequences without requiring protected attribute annotations, thereby extending robustness from static to dynamic learning scenarios.

Knowledge Transfer and Post-Training Compositionality. It's common to deal with many trained models as monolithic entities whose capabilities cannot be easily (or efficiently) transferred or composed. When a new version of a foundation model is released with improved pre-training, practitioners must either discard previous fine-tuning efforts and retrain from scratch or continue using outdated backbones. Similarly, when multiple specialized models are trained independently, combining their capabilities into a unified system typically requires joint retraining on aggregated data. These limitations impose substantial computational costs and constrain the modularity of AI systems.

This motivates a fundamental question: how can learned knowledge be represented and manipulated to enable efficient transfer and composition across models? Recent work on *task vectors* [82], which encode fine-tuning as additive weight-space directions $\tau = \theta^{ft} - \theta^{pre}$, suggests that model updates can be treated as a vector. However, task vectors are typically architecture-specific, limiting their transferability across models with different initialization or pre-training strategies.

In **Chapter 8**, we investigate when and how task vectors can be transported across models. We introduce *Transfusion*, a permutation-based alignment procedure specifically designed for Transformer architectures that addresses two critical technical challenges: *head contamination* in multi-head attention layers and *residual path handling* in skip connections. This enables data-free, training-free transfer of task vectors from one pre-trained backbone to another, allowing fine-tuning investments to carry forward as foundation models evolve.

Beyond knowledge transfer, we examine knowledge *composition*: the ability to merge multiple independently fine-tuned models into a single multi-task system. In **Chapter 9**, we establish both theoretically and empirically that local flatness in the loss landscape is a key predictor of merge compatibility. We show that pre-training objectives that encourage wide minima facilitate subsequent linear combinations of task vectors without degrading performance. We validate this principle in 3D medical image segmentation, where privacy constraints and annotation scarcity make model merging particularly valuable.

2.1 Organization of the Thesis

This thesis presents the work of the candidate and his collaborators ("*we*" in the following) and addresses the aforementioned limitations through a series of interconnected contributions. The structure of this thesis is organized as follows:

Book II addresses shortcut learning and spurious correlations across static, continual, and multimodal settings:

- In **Chap. 3**, we review the literature on spurious correlations, group-distributional robustness, bias mitigation techniques, continual learning foundations, and psychological tools to introduce the meaning of implicit biases.

- In **Chap. 4**, we introduce ClusterFix, a two-stage framework that leverages self-supervised clustering to identify bias-prone samples without demographic labels. By treating cluster assignment errors as proxies for minority subgroups, we achieve worst-group accuracy improvements that match or exceed methods requiring explicit group annotations [22].

- In **Chap. 5**, we extend robustness considerations to continual learning scenarios, demonstrating that standard rehearsal mechanisms amplify spurious correlations over time. We propose Learning without Shortcuts (LwS), which combines unsupervised clustering objectives with loss-based buffer sampling to maintain worst-group performance across sequential tasks [24].

- In **Chap. 6**, we investigate implicit biases in vision-language models through the lens of social psychology, adapting the Stereotype Content Model and Implicit Association Test to quantify differential associations in CLIP architectures. We evaluate orthogonal projection debiasing strategies and reveal their limitations when applied to models with varying initial bias magnitudes [23].

Book III investigates parameter-space alignment for knowledge transfer and model merging:

- In **Chap. 7**, we review the literature on weight-space permutations and mode connectivity, establishing the theoretical foundations for aligning models with permutation symmetries. Additionally, we introduce concepts for both parameter-space alignment and model merging, with an emphasis on flatness and its connection to merge compatibility.

- In **Chap. 8**, we introduce Transfusion, a structured two-level permutation procedure for Transformer architectures that addresses head contamination and residual connection handling. By employing permutation-invariant spectral metrics for inter-head matching and constrained intra-head permutations, we enable data-free transport of task vectors across models with different pre-training strategies, demonstrating successful knowledge transfer between distinct CLIP backbones [151].

- In **Chap. 9**, we establish both theoretically and empirically that local flatness in the loss landscape predicts successful model merging. We introduce a pre-training strategy that enables an effective merging of independently fine-tuned 3D segmentation models. We demonstrate practical applicability in 3D medical imaging segmentation, where privacy constraints and annotation scarcity make model merging challenging [115].

Book IV concludes the thesis with a discussion of open problems, while **Appendix** provides a summary of the contributions of the candidate and a list of publications and activities carried out by the candidate during his PhD.

Book II

Mitigating Spurious Correlations in Deep Neural Networks

3. Background

Modern neural networks deliver impressive performance across a wide range of tasks [45, 183, 69, 174, 43, 144]. Yet, in the presence of distribution shifts, changing environments, or diverse subpopulations, a core difficulty arises: neural networks tend to follow the *Principle of Least Effort* [63], capitalizing on spurious correlations, surface-level regularities that align with target labels but do not reflect genuine causal factors. This shortcut learning manifests in various ways depending on the deployment setting, and each manifestation demands a tailored mitigation approach.

This chapter establishes the theoretical and empirical foundations for the contributions presented in subsequent chapters, organized around three interconnected perspectives on spurious correlations: (i) *offline debiasing*, where bias patterns remain fixed throughout training (Chap. 4); (ii) *continual debiasing*, where spurious correlations evolve across sequential tasks (Chap. 5); and (iii) *implicit biases in multimodal models*, where stereotypical associations emerge from contrastive alignment objectives (Chap. 6). For each perspective, we review the relevant literature, identify key limitations of existing approaches, and contextualize how our contributions address these gaps.

3.1 Spurious Correlations

Artificial intelligence systems can suffer mainly from two distinct sources of bias: **data bias**, arising from spurious correlations in the training distribution, and **algorithmic bias**, stemming from design choices that favor certain patterns over others [63, 77, 124]. The consequences are well-known: small input perturbations or context changes cause dramatic performance drops [10, 152, 173], and minority groups or underrepresented subpopulations suffer disproportionately high error rates [17, 40, 165]. Enhancing model trustworthiness therefore requires moving beyond aggregate accuracy metrics to consider robustness, fairness, and worst-group performance [16, 62, 64, 104, 179].

A significant body of work addresses distribution shift through domain adaptation and generalization techniques [188, 208]. These approaches typically assume access to multiple source distributions during training and aim to learn

representations that transfer to unseen target distributions. Domain randomization has proven effective in robotics applications [178], while meta-learning frameworks enable rapid fine-tuning to new distributions [51, 182]. Adversarial training offers another avenue, explicitly optimizing for worst-case perturbations to induce robust features [173, 204, 159]. While adversarial objectives can align model behavior more closely with human perception, they do not provide formal guarantees of robustness [104].

3.1.1 Debiasing Without Protected Attributes

Traditional debiasing methods such as Distributionally Robust Optimization (DRO) [145] and Group Distributionally Robust Optimization (GDRO) [154] explicitly optimize for worst-group performance by minimizing the maximum loss over predefined subpopulations. While theoretically principled, these supervised approaches require access to protected-group annotations (*e.g.*, demographic attributes), which are often unavailable due to privacy constraints, ethical considerations, or practical limitations in identifying relevant attributes. This has motivated a shift toward unsupervised debiasing methods that infer group structure from the data itself [129, 133, 106].

Among unsupervised approaches, clustering-based methods have gained prominence. George [168] and BPA [164] both leverage clustering to approximate protected groups: they first train a model via empirical risk minimization (ERM), cluster the learned feature space, and then optimize a debiasing objective using cluster assignments as pseudo-labels. However, these methods face critical limitations. George applies a minimax objective over clusters, which proves vulnerable to outliers and can lead to overfitting on noisy cluster assignments, particularly on datasets with complex bias structures like CelebA [164]. BPA addresses this by dynamically reweighting samples based on per-cluster loss, but inherits the fundamental issue that clustering on ERM features may already reflect spurious correlations present in the initial training.

The work presented in Chap. 4 addresses these limitations by proposing a cluster-based debiasing strategy [22] achieving state-of-the-art worst-group accuracy without protected-group supervision.

3.2 Spurious Correlations in Continual Learning

The debiasing approaches discussed above operate in static settings where the complete training distribution is available from the outset. However, many practical applications require systems that learn from continuous streams of non-i.i.d. data while preserving previously acquired knowledge, a challenge addressed by *continual learning* (CL) [137, 41]. This paradigm introduces the well-known pitfall of **catastrophic forgetting** [121, 58]: neural networks dramatically lose performance on earlier tasks when trained on new ones.

3.2.1 Continual Learning Scenarios

The CL literature defines three primary scenarios [181]: **Task-Incremental Learning** (task identifiers available at inference), **Domain-Incremental Learning** (fixed labels, shifting input distributions), and **Class-Incremental Learning** (Class-IL), where models train on sequential tasks introducing new classes without access to task identifiers at test time [14, 75, 193]. Class-IL is the most widely adopted benchmark as it closely mirrors real-world deployment scenarios [181, 49, 4].

Rehearsal-based methods have emerged as the dominant paradigm [18, 149, 75, 20, 4], maintaining small memory buffers that store representative samples from past tasks and interleave them with new data during training. The effectiveness of rehearsal stems from its ability to provide gradient diversity that approximates the statistical properties of the original data distribution [49, 31]. This strategy has proven remarkably flexible, adapting to complex scenarios including annotation noise [7, 94, 126], partial supervision [15, 92], and absence of task boundaries [18, 41]. Alternative approaches such as **regularization-based methods** [95, 3, 203, 143], **architectural solutions** [118, 153], and **prompt-based methods** [190, 189, 167], generally underperform rehearsal in Class-IL [18, 181].

3.2.2 The Interaction Between Bias and Forgetting

Recent investigations reveal that spurious correlations interact with continual learning in complex ways [156, 102]. When bias patterns evolve across tasks, continual learners face a dual challenge: retaining discriminative knowledge of past classes while maintaining robustness to spurious features whose relevance may shift over time. Empirical evidence shows that transfer bias is exacerbated in CL, affecting both future and past tasks bidirectionally [156, 102].

Current approaches to this problem build upon established rehearsal frameworks. The Balanced Greedy Sampler (BGS) [102] adapts GDRO to continual settings by incorporating group-aware sampling and post-hoc classifier adjustments at task boundaries. While effective as a proof of concept, BGS inherits the limitation of requiring protected-group labels. Learning without Prejudices (LwP) [86] decouples feature extraction from classification by training the latter on synthetically generated data. However, maintaining a generative model on non-stationary data streams introduces additional challenges [138, 180, 60], limiting the method's applicability.

The work presented in Chap. 5 addresses these limitations by proposing continual debiasing strategies that combine unsupervised bias detection with rehearsal mechanisms designed to preserve both accuracy and fairness across evolving task sequences [24]. By jointly optimizing for balanced replay on buffer data and debiasing regularization on streaming data, our approach ensures robustness while avoiding catastrophic forgetting.

3.3 Implicit Biases in Vision-Language Models

The preceding sections examined explicit spurious correlations and observable disparities in model performance across identifiable subpopulations, as measured by group-level accuracy metrics. However, a more subtle challenge arises for foundational vision-language models such as CLIP [144], which have revolutionized tasks including retrieval [67] and recognition [184]. Pre-trained on massive internet-scale datasets, these models inherit not only broad visual-semantic knowledge but also implicit biases embedded in their training corpora. Unlike explicit spurious correlations, implicit biases operate at a more nuanced level, shaping model behavior through stereotypical associations that may not surface in standard benchmark evaluations [5].

3.3.1 The Nature of Implicit Biases

Biases in vision-language models often stem from *spurious correlations* in pre-training data, where models inadvertently associate unrelated attributes, potentially leading to biased decisions and unfair outcomes in deployment [62, 64]. The consequences extend far beyond academic benchmarks. Indeed, biased AI models can lead to disproportionately impacting marginalized communities in healthcare [53, 157, 141, 134], in criminal justice and employment [132].

Social psychology offers crucial insights through its long-standing distinction between *implicit* and *explicit* biases [8, 65]. While explicit biases correspond to conscious attitudes that individuals can report directly, implicit biases operate automatically and unconsciously, influencing behavior even when individuals consciously reject prejudiced beliefs. Despite various proposed debiasing strategies, from supervised methods that adjust training data based on protected attributes [155, 86] to unsupervised techniques that modify training objectives [22, 164, 206, 129], standard bias evaluation benchmarks usually fail to evaluate the complex interplay of hidden biases that influence model outputs [5].

3.3.2 Psychological Frameworks for Bias Measurement

Drawing on multidisciplinary perspectives, we adopt three established psychological frameworks to assess implicit biases in vision-language models. These instruments reveal hidden patterns of discrimination that conventional accuracy-based metrics cannot detect.

Stereotype Content Model (SCM). The Stereotype Content Model [52] posits that social group perception organizes along two fundamental dimensions: *Competence* and *Warmth*. These dimensions capture complementary aspects of social cognition—whether a group is perceived as capable of enacting its intentions (Competence: Intelligent, Competent, Skillful) and whether those intentions are benevolent or harmful (Warmth: Friendly, Warm, Likable). The SCM framework has proven universal across cultures and contexts Tab. 3.1.

Table 3.1: Overview of models and attributes used to measure implicit biases.

Model	Attributes
SCM	**Competence**: Competent, Intelligent, Skillfull
	Warmth: Warm, Friendly, Likeable
Emotions	**Pos**: Surprise, Attraction, Pleasure, Compassion, Serene, Happiness
	Neg: Anger, Disgust, Fear, Shame, Bitterness, Contempt
Semantic	**Pos**: Positive, Warm, Trusting, Friendly, Respectful, Admirable
	Neg: Negative, Cold, Suspicious, Hostile, Contemptive, Disgusting

Emotions Attribution. Intergroup emotions theory recognizes that emotional responses toward social groups shape downstream behaviors [68, 171]. Research demonstrates asymmetric emotion attribution: individuals tend to experience negative emotions toward outgroup members (fostering prejudice, discrimination, and conflict) while positive emotions toward ingroup members promote cohesion and identity [9, 76]. For AI systems deployed in socially sensitive contexts, the capacity to recognize and avoid perpetuating harmful emotional associations becomes critical Tab. 3.1.

Semantic Differential Scale. The Semantic Differential Scale [136] provides a method for measuring the connotative meaning of concepts through bipolar adjective pairs. This technique reveals the affective coloring that models assign to different demographic groups, exposing subtle evaluative biases along dimensions including warm-cold, trusting-suspicious, friendly-hostile, respectful-contemptive, and admirable-disgusting Tab. 3.1.

3.3.3 Existing Debiasing Approaches

Various methods have been proposed to address these biases by using balanced data during training [116, 42]. Novel approaches involve debiasing vision-language models by projecting out biased directions in text embeddings using adversarial prompts [35]. While effectively reducing some explicit biases, these methods do not adequately address implicit ones [5], highlighting a fundamental gap in current debiasing strategies.

The work presented in Chap. 6 addresses this limitation by adapting psychological measurement frameworks (Implicit Association Test, Common Language Effect Size) to quantify implicit biases across 90 Open-CLIP models and proposing SCM-guided orthogonal projection techniques that calibrate debiasing using social psychology attributes rather than arbitrary visual features [23]. By grounding bias measurement and mitigation in established psychological constructs, our approach reveals discrimination patterns that standard fairness benchmarks overlook and provides a principled framework for post-hoc debiasing without retraining.

4. Overcoming Shortcut Learning via Robust Optimization

4.1 Overview

There has been widespread interest in debiasing methods that aim to mitigate unintended solutions. Debiasing interventions can occur before the learning procedure (pre-processing), during model training (in-processing), or after training (post-processing)[97]. In particular, in-processing approaches directly affect algorithm design and effectively mitigate biases. Proposed methods directly debias the model adjusting sample importance [96, 154, 169] and using adversarial learning [26, 105]. Other techniques employ quantitative fairness metrics as regularization terms [25] and optimization constraints [78]. Although these works address the problem, they require prior knowledge of protected groups, achieved by grouping samples according to their target and bias attribute values, such as gender. Accessing such information is often infeasible due to privacy and ethical constraints. Furthermore, identifying and quantifying bias attributes a priori can pose challenges in complex systems (e.g., organism, climate, and cognition).

So, what if the protected groups are missing during the learning phase?

Our goal is to train a model to mitigate performance disparities among protected groups without exploiting such information during training. We aim to achieve this objective while ensuring satisfactory average performance. These qualities are particularly essential for practical decision-making algorithms. As an example, in healthcare and forensic genetics, inaccurate outcomes can pose significant risks to underrepresented communities, calling for fair and robust solutions [32, 119, 133, 142]. An approximation is needed whenever the system does not access protected groups directly.

Our hypothesis posits that classification of cluster assignments can leverage

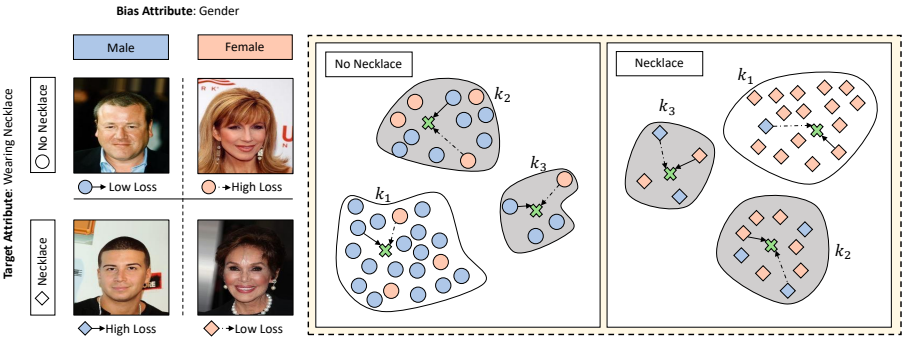


Figure 4.1: An illustrative scenario depicting dataset partitions biased towards gender attributes. We propose that challenging clusters for classification may leverage significant samples that do not possess the same protected attribute. We emphasize the need to pay closer attention to such distributions to address and mitigate model shortcuts.

shared features between bias-aligned and unbiased samples, thereby providing an opportunity to mitigate spurious correlations effectively. Similar to the works of George [168] and BPA [164], our objective is to address this issue by estimating protected groups by clustering.

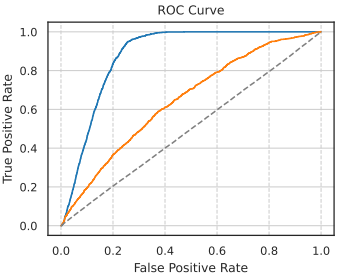
Existing approaches often encounter a common challenge in the presence of dataset bias: overfitting to noisy, densely populated clusters identified during *Empirical Risk Minimization* (ERM) pre-training, where the focus is solely on minimizing the training loss for the target variable y [164]. Furthermore, these methods rely heavily on the assumption that critical samples can be easily identified by clustering in the ERM feature space. However, real-world scenarios do not necessarily support this assumption, as critical samples may be dispersed across clusters rather than concentrated in a single cluster.

In contrast, our approach takes a different standpoint by acknowledging the dispersed distribution of critical samples within the feature space. Our method involves clustering a self-supervised feature space, but, in addition to optimizing the target objective, we also seek to minimize the cluster classification error. Consequently, we define our weighting mechanism policy by incorporating the auxiliary clustering loss with the target objective. To demonstrate the effectiveness of our approach, we provide a compelling illustration in Fig. 4.1.

Identifying Bias. As illustrated in Fig. 4.1, we consider a dataset partitioned on a target label, such as "Wearing Necklace", where each partition is biased towards a protected attribute like gender. Following the clustering process and assuming these assignments as ground truth, a classifier may encounter challenges within a cluster where the majority of individuals x do not possess the

same protected attribute (highlighted in Fig. 4.1 as critical samples within the grey clusters). To address this challenge, we aim to learn new features that can identify a common per-cluster attribute and achieve satisfactory performance. This means looking for features that are consistent with the original cluster assignment, but different from the protected attribute. We believe that upweighting samples from clusters with high cluster classification loss can help identify data distributions that enhance model robustness and mitigate bias.

Observation. To validate our intuitions, we conducted an experiment on the CelebA dataset (denoted as D) with the target attribute y representing Wearing Necklace and the protected attribute a representing gender (female $= 0$ and male $= 1$). In addition, we define a group g as $g = (y, a)$. The target label exhibits a strong correlation with female individuals: within the training set D_t, comprising a total of $162,770$ samples, $D_{y=1}$ consists of $18,525$ females and only $1,239$ males.



As aforementioned, we defined critical groups as $g_0 : (y = 0, a = 0)$ and $g_1 : (y = 1, a = 1)$. Consequently, the label z define if $x \in \{g_0, g_1\}$ or not. When training an ERM classifier on y, we observed that the accuracy for the group g_1 (male with necklace) was only 2.72%. Subsequently, based on self-supervised features, we obtained cluster assignments for each partition D_y using k-means with $k = 8$. Afterward, we trained an ERM cluster classifier using these assignments. The inset figure presents the ROC curve for partition $D_{a=1}$ (blue) and partition $D_{y=1}$ (orange), which quantifies the correlation between the critical label z and the ERM cluster-objective. Interestingly, we observed a correlation between cluster errors and critical samples in these partitions, indicating the former can be adopted profitably to design an ad-hoc weighting strategy, which is the main proposal of our work. In Sec. 4.2.2, we will formally define how our policy detects critical distributions during the learning phase.

4.1.1 Problem Setup

We are given a set of n datapoints $x_1, ..., x_n \in \mathcal{X}$ associated with a binary label $y_i \in \{0, 1\}$. In addition, each datapoint is associated with a label $c_i \in \{1...C\}$ defined by cluster assignment. We train a deep neural network composed of a feature extractor $\mathcal{F} : X \to R^d$, a binary classifier $\mathcal{T} : R^d \to R^2$, and a set of auxiliary classifiers $\mathcal{C}_y : R^d \to R^C$, one for each task label y. The classification performance of y is evaluated on pre-defined groups G based on the average and worst-group accuracy as in [154]. The group label $g \in \mathcal{G}$ is only used for metric evaluation and is not accessible during optimization. The objective is to

achieve good average and worst-group accuracy at test time without training group annotations.

4.1.2 Relevant Objective Functions

In this section, we describe in detail four training approaches to introduce our work in the next section.

Empirical Risk Minimization. Typically, a model defined by a feature extractor $\mathcal{F}$ and a target classifier $\mathcal{T}$ try to solve the following optimization problem:

$$\arg\min_{\mathcal{F},\mathcal{T}} \frac{1}{N} \sum_{i=1}^{N} \ell(\mathcal{T} \circ \mathcal{F}(x_i), y_i) \tag{4.1}$$

where ℓ represents a loss function, y is the target, and x is the input. In other words, ERM minimizes the average loss across data points. In general, this kind of procedure needs better generalization in some groups during inference.

Group Distributionally Robust Optimization [154]. The Group DRO approach utilizes training group annotations to reduce the maximum group error within the training set. Assuming the availability of group annotations for the training data, the objective function for Group DRO can be expressed as:

$$\arg\min_{\mathcal{F},\mathcal{T}} \max_{g \in \mathcal{G}} \left\{ \frac{1}{N_g} \sum_{i=1}^{N_g} 1(g_i = g) \ell(\mathcal{T} \circ \mathcal{F}(x_i), y_i) \right\} \tag{4.2}$$

George [168]. Since g is often not available in real scenarios, George approximates protected groups with cluster assignments. More specifically, this method is organized as follows: (*i*) train a model via ERM, (*ii*) cluster the feature space, (*iii*) train the final model from scratch with clustering information as in Eq. 4.3. The central empirical hypothesis here is that the feature space of deep neural networks trained via ERM carries information about group labels.

$$\arg\min_{\mathcal{F},\mathcal{T}} \max_{c \in \mathcal{C}} \left\{ \frac{1}{N_c} \sum_{i=1}^{N_c} 1(c_i = c) \ell(\mathcal{T} \circ \mathcal{F}(x_i), y_i) \right\} \tag{4.3}$$

BPA [164]. BPA takes advantage of the same hypothesis of George: protected groups are approximated by clustering after ERM pretraining. On the other hand, the optimization process is guided by dynamically reweighting sample importance.

$$\arg\min_{\mathcal{F},\mathcal{T}} \frac{1}{N} \sum_{i=1}^{N} w_{c_i} \ell(\mathcal{T} \circ \mathcal{F}(x_i), y_i) \tag{4.4}$$

where the sample weight w_{c_i} is given by the following equation:

$$w_c = \frac{1}{N_c}\mathbb{E}_{(x,y)\sim P_c}[\ell(\mathcal{T}\circ\mathcal{F}(x_i), y_i)] \tag{4.5}$$

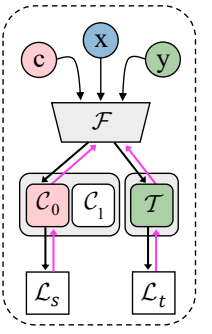
ClusterFix. This paper presents ClusterFix (CFix), a practical in-processing debiasing framework aiming to jointly optimize a classification task and an auxiliary one to generalize over groups without their supervision. Our evaluation demonstrates that the proposed solution outperformed the previous state-of-the-art unsupervised debiasing approaches. Moreover, despite not having group annotations during training, ClusterFix outperformed the supervised approach GDRO [154], even without explicit bias information.

Our contributions can be summarized as follow: (1) We propose an effective debiasing method to mitigate inductive bias in real datasets, which does not rely on prior knowledge on protected groups; (2) We show how cluster assignment classification is a practical auxiliary task that improves worst-case robustness and generalization; (3) Our approach reaches state-of-the-art performances in standard bias-removal benchmarks, even w.r.t. supervised methods.

4.2 Proposed Method: ClusterFix

ClusterFix (CFix) is a two-stage debiasing approach that does not require protected group supervision. Similarly to previous works, CFix weighs the contribution of each example to the overall classification loss as in Eq. (4.4); differently, the importance of each example does not depend only on the mismatch between the prediction and ground-truth label y, but also on an additional term that considers how examples cluster in latent space. Briefly, our approach consists of two subsequent steps:

Cluster Assignment. The first stage regards preparing a pretext task, which will be optimized in the second stage. In particular, the idea is to cluster the latent space of a deep neural network into several groups, thus yielding novel pseudo-labels c from the computed assignments to the clusters. In doing so, CFix leverages self-supervised pre-training (performed with Barlow Twins[202]), whose latent space is subsequently clustered using the k-means algorithm. Such a preliminary stage reduces the risk of biased representation in the downstream task. Indeed, the pseudo-labels produced in this way are opaque with respect to the target bias-prone task, as they rely only on self-extracted features.



Debiasing Training. In the second stage, we ask the model to pursue a two-fold objective: not only it has to learn the target task embodied by y, but there is a tailored objective that constrains the feature space. In particular, it seeks to ensure the model remains consistent with the original cluster assignments c.Eventually, to preserve minority groups, we weight each example proportionally to the average *error* of its cluster, where *error* is defined as the difference between the original and pseudo labels y and c. In this respect, what differentiates CFix from existing approaches is that the clustering membership actively and directly contributes to the overall learning objective.

As an example, George [168] also performs a preliminary clustering step; however, while its authors exploited clustering to compute the classification loss w.r.t. an independent sample grouping (represented by c), we instead use a metric related to c to weight examples.

To provide an intuition, we refer to the observation in Sec. 4.1 in which a clear dependence arises between an independent cluster's assignment metric (e.g., an ERM pre-trained classifier) and the presence of protected features. For this purpose, this suggests that the empirical risk incurred by a wrong cluster assignment can be used as a proxy for the importance of an element in terms of its contribution to the task loss on y.

4.2.1 Step 1: Cluster Assignment

First, the dataset was divided by label, from which features were extracted from a pre-trained model. Next, k-means was applied to obtain cluster assignments c, which were used as categorical labels for the next stage. Specifically, a self-supervised-trained model $\mathcal{F} : X \to R^d$ was used for feature extraction.

4.2.2 Step 2: Debiased Training

Let $\mathcal{F} : X \to R^d$, $\mathcal{T} : R^d \to R^2$, and $\mathcal{C}_y : R^d \to R^C$ represent the functions of the pre-trained encoder, task classifier, and cluster classifier respectively, Fig. 5.2. The objective function of the proposed method consists of two parts: a weighted classification loss and a cluster-structural loss. In the following, we provide details on these two separate terms and the re-weighting strategy for the proposed model.

Clustering-Structural Loss. Achieving smoothness in the feature space through cluster classification can prevent model shortcuts in y classification and mitigate inductive bias. This method is also valuable for identifying problematic clusters with high average entropy and small size. We hypothesize that there is a correlation between high-entropy clusters and out-of-distribution elements, which can help the model generalize better in worst-case scenarios. This approach's benefits are underpinned by information theory [197], as the structural loss creates

an information bottleneck that facilitates the identification of proxy objective bias and improves model generalization. Formally, the clustering-structural loss is defined as follows:

$$\mathcal{L}_s = \sum_{i=1}^{N} \ell(\mathcal{C} \circ \mathcal{F}(x_i), c_i) \tag{4.6}$$

Task-Weighted Classification Loss. The main objective of the optimization process is to learn how to classify y, which is achieved through task-weighted classification loss. Each sample's classification loss is weighted by a factor w_k, reflecting the significance of the cluster to which it belongs. The weight is determined based on the average loss with respect to y and the structural loss with respect to the clusters. The task-weighted classification loss is defined as follows:

$$\mathcal{L}_t = w_{c_i} \ell(\mathcal{T} \circ \mathcal{F}(x_i), y_i) \tag{4.7}$$

where w_c is:

$$w_c = \frac{1}{N_c} \mathbb{E}_{(x,y)\sim P_c}[\ell(\mathcal{T} \circ \mathcal{F}(x_i), y_i) + \gamma \ell(\mathcal{C}_{y_i} \circ \mathcal{F}(x_i), c_i)] \tag{4.8}$$

ClusterFix Objective. In summary, the proposed debiasing method's overall objective function can be expressed as the following optimization problem, the trade-off hyper-parameter is denoted by γ:

$$\min \mathcal{L}_t + \gamma \mathcal{L}_s \tag{4.9}$$

4.3 Experiments

In this section, we present the experiments to evaluate the proposed debiasing method and compare it with several state-of-the-art debiasing techniques. We used the same experimental settings as BPA [164] to ensure a fair comparison among different debiasing methods. Specifically, we trained the proposed method using a ResNet-18 as the backbone architecture with the Adam optimizer, a learning rate of 1×10^{-4}, a batch size of 256 images, and a weight decay rate of 0.01 for 100 epochs. The learning rate was scheduled with cosine annealing. For all experiments, we performed k-means clustering with $K = 8$ and the cluster weight of the c_{th} cluster, ω_c, was updated with a momentum $m = 0.3$ as in [164]. Additional experiments on hyperparameters and datasets are available in the supplementary material.

4.3.1 Benchmarks

CelebA. CelebA [108] is a dataset comprising 202,599 celebrity face images with 40 binary attribute annotations for each image. Moreover, in our experiments, the gender attribute has been used as the bias attribute to evaluate the

Table 4.1: Unbiased accuracy (%) on CelebA dataset. **Ours** denotes ClusterFix. Best unsupervised results in **bold**.

	Unsupervised				Supervised
Target	**ERM**	**George** [168]	**BPA** [164]	**Ours**	**GDRO** [154]
Double Chin	64.61± 0.82	76.23± 0.11	82.92± 0.54	**85.13**± 0.30	83.19± 1.11
Pale Skin	71.50± 1.60	78.22± 3.75	90.06± 0.75	**91.17**± 0.04	90.55± 0.84
Chubby	67.42± 0.95	74.88± 1.91	83.88± 0.36	**84.16**± 0.22	81.90± 0.20
Necklace	55.04± 0.59	58.79± 0.10	68.96± 0.12	**68.99**± 1.19	62.89± 3.69
Wearing Hat	93.53± 0.37	95.72± 0.71	96.80± 0.26	**97.88**± 0.09	96.84± 0.46
Big Lips	60.87± 0.58	64.99± 0.13	**66.50**± 0.24	65.40± 0.48	63.70± 0.44
Bangs	89.04± 0.47	92.62± 0.12	93.94± 0.57	**94.67**± 0.16	94.45± 0.17
Hairline	69.72± 0.78	78.86± 0.40	84.95± 0.49	**87.00**± 0.12	85.15± 1.31
Wavy Hair	73.10± 0.56	77.39± 0.15	**79.89**± 0.71	79.42± 0.12	79.65± 0.63
Brown Hair	78.07± 0.87	83.07± 0.07	83.83± 0.66	**85.30**± 0.47	84.87± 0.07
Average	72.29	78.07	83.17	**83.91**	82.31

robustness of the proposed method, as in [164]. We initialized the feature extractor $\mathcal{F}$ parameters by using a self-supervised pre-trained network with Barlow Twins [202]. In particular, for the self-supervised training, we used an output dimension of 1 024, a batch size of 256, and an SGD optimizer with a fixed learning rate of 0.6. We set the λ parameter to 0.5.

Following [164], we focused on gender as the fixed bias attribute and excluded 8 out of 40 attributes due to limited samples in the test set. Among the remaining 32 attributes, 26 exhibited a significant correlation with gender, showing a classification accuracy gap of over 5% compared to unbiased accuracy [154]. To explore diverse scenarios, we selected the top 5 attributes with the highest gap and the bottom 5 attributes with the lowest gap, as identified in [164].

Waterbirds. The Waterbirds dataset [154] is designed to evaluate the robustness of deep networks w.r.t. spurious correlations and distribution shifts. It has been created by selecting images of birds from the Caltech-UCSD Birds-200-2011 dataset [186] and overlaying them on backgrounds obtained from the Places dataset [207]. This dataset includes two attributes: the type of bird (waterbird or landbird) and the background (water or land). The training set comprises 4, 791 samples. 56 of the 1, 113 waterbird samples have a land background, while 180 of the 3, 678 landbird samples have a water background. To assess model robustness, the validation and test sets evenly distribute landbirds and waterbirds across land and water backgrounds. We initialized $\mathcal{F}$ with ResNet18 pre-trained on ImageNet and set the λ parameter to 0.01.

Evaluation Protocol. To evaluate the proposed method, we calculate the accuracy of every group $g = (y, b)$, defined as a combination of target and

Table 4.2: Worst-Group accuracy (%) on CelebA dataset.

	Unsupervised				Supervised
Target	**ERM**	**George** [168]	**BPA** [164]	**Ours**	**GDRO** [154]
Double Chin	21.33± 2.24	50.00± 0.41	67.78± 0.91	**74.26**± 3.94	72.94± 1.14
Pale Skin	36.64± 3.53	62.03± 1.65	**88.60**± 1.48	87.01± 1.46	87.68± 2.37
Chubby	24.30± 3.73	58.01± 1.10	72.32± 0.93	71.01± 1.17	**72.64**± 1.70
Necklace	2.72± 0.83	13.82± 0.41	41.93± 2.47	**55.56**± 0.38	24.34± 7.81
Wearing Hat	85.12± 0.31	92.93± 0.76	94.94± 0.19	**96.58**± 0.63	94.67± 0.41
Big Lips	30.85± 0.62	44.51± 0.83	56.99± 3.05	**57.27**± 0.58	47.55± 1.03
Bangs	76.91± 3.27	85.90± 0.24	92.21± 1.24	**93.01**± 0.36	92.12± 1.03
Hairline	35.69± 0.35	57.30± 0.90	79.11± 1.91	**84.15**± 0.82	79.12± 2.11
Wavy Hair	38.01± 0.85	53.17± 0.43	65.74± 1.13	**69.92**± 0.38	66.79± 1.62
Brown Hair	59.58± 2.55	73.20± 0.88	71.50± 0.97	**79.18**± 0.50	78.92± 1.61
Average	41.11	59.09	73.11	**76.80**	71.67

Table 4.3: Unbiased and Worst-Group accuracy (%) on the Waterbirds dataset.

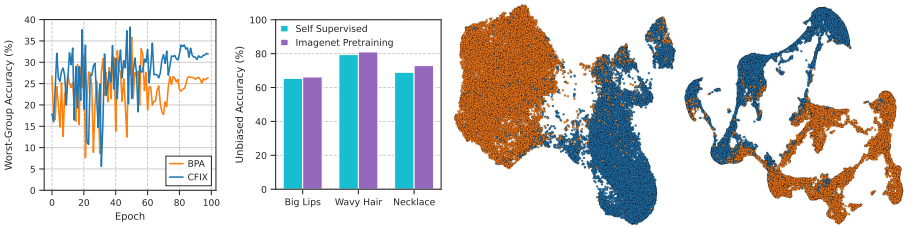
Unbiased Accuracy (%)					
		Unsupervised			Supervised
Target	**Bias**	**ERM**	**BPA** [164]	**Ours**	**GDRO** [154]
Object	Place	84.63± 0.00	87.05± 0.00	**86.29**± 0.00	88.99± 0.00
Place	Object	87.99± 0.00	88.44± 0.00	**92.17**± 0.00	89.20± 0.00
Worst-Group Accuracy (%)					
Target	**Bias**	**ERM**	**BPA** [164]	**Ours**	**GDRO** [154]
Object	Place	62.39± 0.00	71.39± 0.00	**74.03**± 0.00	80.82± 0.00
Place	Object	73.34± 0.00	79.16± 0.00	**86.61**± 0.00	85.27± 0.00

bias attribute values. The bias attribute $b \in \{1...B\}$ is an external annotation unused during optimization (e.g., gender, background). We report results in terms of average-group accuracy (**unbiased accuracy**) and **worst-group accuracy** [154, 164]. All reported results are the average of three runs.

4.3.2 Comparisons with Other Debiasing Approaches

The proposed method is compared with several state-of-the-art debiasing techniques:

- **Empirical Risk Minimization (ERM)**: our vanilla baseline. It trains a model on a biased dataset, leading to a biased model;
- **Learning from Failure (LfF)** [129]: it trains a debiased classifier using the misclassification information of the biased classifier;



(a) (b) (c) (d)

Figure 4.2: Robustness evaluation of models trained without critical samples a and the effectiveness of CFix backbone pre-training on unbiased accuracy measurement b. c and d are UMAP projections visualizing CFix feature space (*Wearing Necklace* = false) at the initial and final epoch. Blue and orange colors represent male and female gender values, respectively.

- **George** [168]: a debiasing method that approximates bias attributes with cluster assignment and weights the objective function to maximize the worst-case group accuracy;
- **Debiased Representations with Pseudo-Attributes (BPA)** [164]: a debiased representation is learned by introducing cluster-generated pseudo-attributes;
- **Group Distributionally Robust Optimization (GDRO)** [154]: this method optimizes the worst-case performance over a distributionally robust uncertainty set using explicit bias supervision.

Main Results. The results of our evaluation on CelebA are presented in Tab. 4.1 and Tab. 4.2. We show how CFix outperforms all competitors across all evaluated scenarios, achieving higher unbiased and worst-group accuracy metrics. Specifically, our method outperforms the bias-supervised approach, Group DRO, and state-of-the-art BPA in both metrics. Additionally, our optimization process demonstrates greater stability across different runs. The improvement in worst-group accuracy is noteworthy, given that other approaches, such as George and Group DRO, prioritize maximizing worst-case group accuracy, which is not the focus of our approach. In more detail, our method achieves an average improvement of +0.74% in average accuracy and +3.27% in worst accuracy (+13.63% for *Wearing Necklace* target) compared to the previous unsupervised state-of-the-art. Additionally, experiments on Waterbirds confirm the effectiveness of our approach even in a controlled environment, as shown in Tab. 4.3. We observed improvements in the worst-case performance for Object (*bird*) and Place (*background*) classification, with gains of +2.64% and +7.45%, respectively.

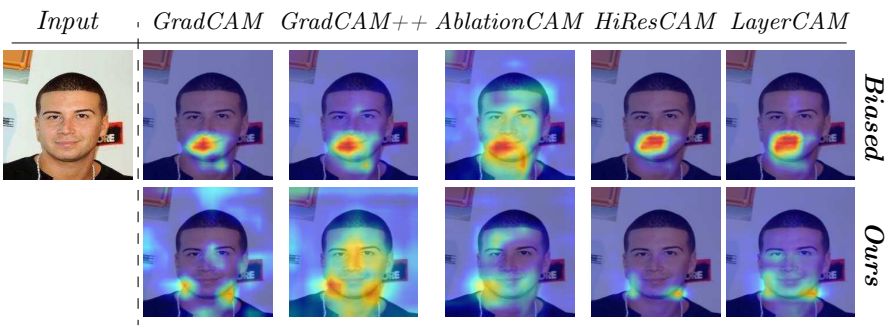


Figure 4.3: Visualization of the class activation maps generated by Grad-CAM [163], Grad-CAM++[28], Ablation-CAM[146], HiResCAM [46], and LayerCam [88] for the ERM and the proposed debiased approach targeting the *Wearing Necklace* attribute.

4.4 Model Analysis

Ratio of the Bias-Aligned Samples. To showcase the resilience of different approaches to distribution shifts, we have designed an experiment in which we selectively remove bias-aligned samples from the training dataset while keeping the test set unchanged. Our study focuses on the Waterbirds dataset, where we remove minority groups from the training set (i.e., all waterbirds on land backgrounds and vice versa). The worst group accuracy on the test set is reported across epochs in Fig.4.2a. Our results indicate that CFix outperforms BPA on minority groups not present in the training distribution. This provides evidence that the proposed structural loss enhances the overall worst-case generalization performance.

On the Effects of Pre-Training. The feature extractor $\mathcal{F}$, used for clustering and backbone initialization, is a crucial component in our pipeline. To verify the effectiveness of our approach on CelebA while altering the backbone, ResNet-18 pre-trained on ImageNet has been employed to initialize $\mathcal{F}$. We conducted identical experiments for the best and worst target attributes as specified in Tab. 4.1. The outcomes illustrated in Fig. 4.2b indicate minimal changes in overall performance, with a slight improvement when using the latter setting. These results suggest that it is possible to leverage bias-aligned signals without requiring an Empirical Risk Minimization (ERM) model for each target. Notably, we employed the same self-supervised encoder $\mathcal{F}$ to perform clustering for all CelebA experiments, while others trained a separate ERM model for each target. Refer to the supplementary material for ERM pre-training experiments.

Feature Space Visualization. In Fig. 4.2c and Fig. 4.2d, we present visualizations of UMAP [123] projections on the CelebA dataset for the Wearing Necklace classification. Specifically, we visualize only negative examples (*Wear-*

ing Necklace = false) to effectively show the feature space at the initial and final epoch using CFix. Our observations suggest that our proposed model successfully achieves a smoother feature space within the same class, mitigating the presence of large clusters caused by shortcut solutions. Other methods aim to mix the bias attributes in the feature space to improve worst-case generalization. In contrast, our findings advocate that good worst-group and average generalization properties can be achieved even when the bias attributes are separable in the feature space. Therefore, *unlearning* the bias attribute is not always necessary to avoid shortcuts and achieve good model performance.

Model Explainability. The experiments in Fig. 4.3 explore the interpretability of the proposed debiased approach compared to the classical Empirical Risk Minimization (ERM) method. Specifically, we use several explainability techniques to visualize the class activation maps for the *Wearing Necklace* target on CelebA. For instance, the ERM method emphasizes features that are not correlated with the target attribute, while CFix prioritizes regions that are more indicative of the regions of interest. These results suggest that our method mitigates unintended solutions.

4.5 Discussion

Mitigating model shortcuts without directly observing bias attributes is a challenging, relatively unexplored task for achieving bias removal in deep networks. Previous research has attempted to address this problem by modifying the target objective using pseudo-groups identified through ERM pre-training. However, our empirical findings indicate that this approach could be suboptimal in some scenarios. Our key contribution is the recognition that empirical cluster error can serve as a proxy for identifying samples likely to be affected by the inductive bias of deep networks. By leveraging this insight, CFix effectively up-weights such samples, improving the worst-case and average generalization for protected groups across multiple standard benchmarks. Our study demonstrates that ClusterFix and the insights gained from experimental results offer a robust foundation for advancing worst-case generalization and algorithmic fairness without relying on demographic data.

5. Avoiding Forgetting in the Presence of Spurious Correlations

5.1 Overview

Modern AI systems are trained on an ever-increasing volume of data, much of which may not be available during the initial training phase, *e.g.* new tasks or classes can be discovered as the system evolves. For this purpose, **Continual Learning** (CL) has become a prominent paradigm, especially when privacy concerns or limited resources constrain access to previous data. In CL, models learn tasks sequentially, facing the challenge of mitigating *catastrophic forgetting* [121, 137], where the model forgets previously acquired knowledge while learning new tasks. In this respect, numerous CL methods exploit a *rehearsal* mechanism to protect against forgetting [19, 111, 4, 18, 13]. These methods utilize a small memory buffer to store past data and alternate training between the current task and the examples stored within the buffer. The sampling strategy typically employed to add or remove examples is *reservoir sampling* [150, 185], a stochastic method that ensures equal representation of previous tasks within the buffer.

Due to its broad applicability, the intersection between bias-related issues and CL has been recently studied in [86]. We build on this research line, arguing that rehearsal methods have significant limitations when applied to tasks influenced by bias and spurious correlations. Since the memory buffer holds only a small, random subset of past examples, it is likely to be dominated by instances that exhibit spurious correlations, potentially leading to underrepresented groups being unfairly penalized. As the buffer samples are the only source of insight into past tasks, a buffer poisoned by spurious correlations could further amplify existing biases, creating a compounding effect.

To illustrate the issue, we direct the attention of the reader to Fig. 5.1. In CelebA [110], attributes like *Wearing Necklace* exhibit strong correlations with

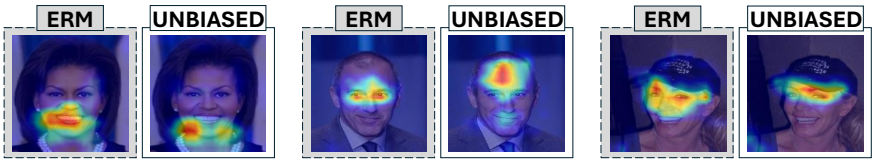


Figure 5.1: Attention heatmaps for *Wearing Necklace*, *Receding Hairline*, and *Wearing Hat* attributes in CelebA, using Empirical Risk Minimization (ERM) and ClusterFix [22]. ERM models often concentrate on irrelevant regions, exploiting shortcuts, whereas CFIX focuses on more salient features.

latent variables like *Gender*, *i.e.*, wearing a necklace is more common among women. As a result, the model is prone to learning shortcuts [62], associating the presence of a necklace with female traits and its absence with male traits. Such a shortcut can lead the model to predict a necklace on a woman even when she is not wearing one, or, conversely, fail to recognize a necklace on a man who is. To avoid shortcuts, current *debiasing methods* [154] exploit expensive auxiliary annotated metadata (*e.g.*, gender or ethnicity), or training paradigms whose outputs are invariant to biases [107]. However, none of these approaches were designed to handle a continuous stream of potentially biased tasks.

So, how do spurious correlations affect rehearsal-based methods?

In light of these intuitions, in Fig. 5.2 we propose a novel approach, **Learning without Shortcuts** (**LwS**), to mitigate the effects of spurious correlations in CL without relying on latent-variable supervision. LwS introduces *i)* an unsupervised objective against shortcuts while training on the current task, and *ii)* a loss-based sampling algorithm to ensure a fair representation across the groups in the buffer population. We conducted experiments on three benchmarks and achieved notable improvements in average and worst-group accuracy, with our results sometimes surpassing methods that employ latent supervision.

5.2 Problem Definition

Spurious Correlations. AI methods often focus on the interaction between an input space, represented as X (*e.g.* an image), and its associated output space, Y (e.g., ground truth label). In this context, we introduce the latent variable z. This variable captures a unique attribute of an element $x \in X$, ranging from broader aspects, such as the presence of artifacts, to more detailed image features, such as the green grass in the background. To define this concept precisely, we can describe an element x with a set of binary attributes $A = \{z_1, z_2, \ldots, z_n\}$.

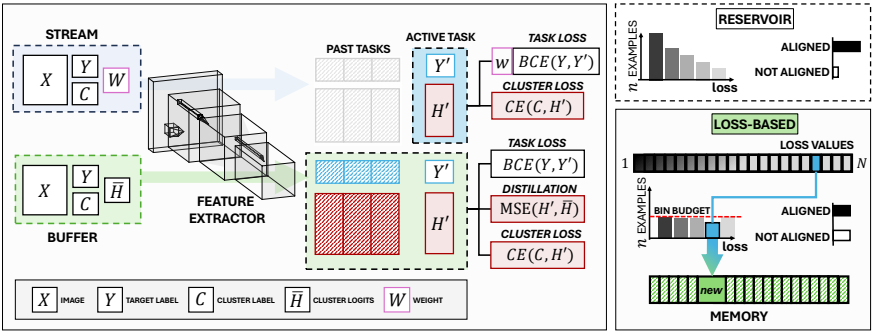


Figure 5.2: Overview of the proposed framework called Learning without Shortcuts (Lws). *Left*: during training, Lws employs tailored optimization objectives to relieve both forgetting and shortcut learning. Specifically, Lws couples the standard cross-entropy loss on labels Y with an auxiliary self-supervised term (*i.e.*, cluster loss). Importantly, the training loss for each example is dynamically adjusted to amplify the contribution of underrepresented groups (e.g., women who do not wear a necklace) during training. *Right*: a visual of the loss-based criterion used to insert new elements within the buffer. The loss values serve as a proxy for distinguishing between bias-aligned and unaligned examples, a feature we leverage to achieve balanced representation across groups.

Even though these attributes may correlate with Y, they do not necessarily correspond to an attribute of interest. For example, the presence of a cow (Y) might be correlated with a background of green grass, where $z = 0$ indicates no green grass and $z = 1$ indicates the presence of green grass. While this correlation exists, relying on it can lead to harmful shortcuts in learning: recognizing the presence of grass may be easier, but it does not indicate the presence of a cow. This discrepancy is often referred to as spurious correlations [63]: associations in the data do not imply a causal relationship with the outcome.

Continual Learning with Spurious Correlations. In an incremental setup, the model is trained sequentially on different datasets $D_1, \ldots, D_T$, where each $D_t = (X_t, Y_t)$ represents a supervised classification task. Each dataset introduces some variation compared to the others, making the tasks distinct. For example, each task could involve classifying a different visual attribute. The objective is to develop a function $f : X_t \rightarrow Y_t$ that effectively integrates new knowledge from successive tasks without losing performance on previously learned ones.

Within this context, each dataset D_t may be influenced by different biases. Consequently, the presence of spurious correlations has a detrimental effect on CL, especially on those methods that build upon a memory buffer, like replay-based approaches. Indeed, their effectiveness relies heavily on the quality of samples stored in the buffer, with significant degradation as it becomes contaminated by bias.

5.3 Method

We present *Learning without Shortcuts* (Lws), a continual debiasing approach that mitigates the harmful effects of bias on learning from a data stream while preventing catastrophic forgetting. In particular, we exploit an auxiliary self-supervised approach to reduce bias. This approach is popular in offline settings [22, 164, 168, 198] and exploits pseudo-labeling to regularize the latent representation of the model. Specifically, the pseudo-labels are obtained by clustering the latent space with k-means. Notably, this strategy poses technical challenges in continual learning due to the emergence of new tasks and associated cluster sets. To overcome these issues, we introduce the following **two modules**.

Data Stream. We start by extracting cluster assignments for the samples of the current task. These will be used throughout the task to ensure alignment with the initial representation. Here, the primary goal is to minimize the distance among samples that belong to the same inferred group (cluster) yet share the same class, thereby reducing the mutual information between spurious correlations and target labels within the data stream [168, 164]. This auxiliary task has also been shown to enhance the smoothness of the latent space [198], a property that facilitates the reuse and transfer of features across tasks [13, 50, 162].

Memory Buffer. To address the shortcomings of traditional replay-based methods, we propose a loss-based strategy to update the memory buffer. Specifically, the magnitude of the loss value is utilized to select which examples to store in the buffer. Secondly, we build upon knowledge distillation [73] to form the replay regularization objective. In contrast to common techniques, our method uses the output of the cluster classifier as the teaching signal for knowledge preservation. By doing so, we can maintain cluster coherence across current and future tasks, thereby mitigating forgetting and enriching transfer capabilities.

5.3.1 Data Stream Objective

Cluster Assignment. At the start of each task t, our approach assigns a cluster c to every element within the dataset D_t. This step involves partitioning D_t based on target labels y and performing k-means for each partition with features from a pre-trained frozen model $F_{pre} : X \to R^d$. Notably, the model F_{pre} remains the same across all tasks.

Debiased Training. From the samples of the data stream, our model is given a twofold objective. Firstly, it solves the binary classification task t, where y denotes the ground-truth label. Secondly, it adheres to a specific objective that constrains the feature space. This objective requires the model to remain consistent with the original cluster assignments c.

To ensure that minority groups are not overlooked, we modify the optimization objective to re-weight the importance of each example. In practice, we assign a weight, w_c in Eq. 5.3, proportional to the average *error* and the cardinality of its cluster. The *error* is defined with respect to the original and pseudo labels y and c.

Formally, let $F : X \to R^d$ be the feature extractor and let $\mathcal{T}_t : R^d \to R^1$ and $C_t : R^d \to R^C$ indicate, respectively, the task head and the cluster classifiers. The latter are two linear projections; while the first outputs the logits of the classes of the t-th task, the second is instead relevant for the auxiliary self-supervised objective. The parameters of the feature extractor F and the task head C_t are updated continuously across tasks. Differently, the parameters of the cluster classifier $\mathcal{T}_t$ are optimized only during task t. Formally, the clustering-structural loss is defined as follows:

$$\mathcal{L}_{cluster} = \mathcal{L}_{\text{CE}}(C_t \circ F(x), c) \tag{5.1}$$

The main objective of the optimization process is learning to classify y, which is achieved through a task-weighted classification loss. Indeed, the loss is weighted by a factor w_c, which reflects the "importance" of the cluster to which the sample belongs. Finally, the task-weighted classification loss is defined as follows:

$$\mathcal{L}_{\text{target}} = w_c \mathcal{L}_{\text{BCE}}(\mathcal{T}_t \circ F(x), y) \tag{5.2}$$

where w_c is:

$$\frac{1}{N_c} \mathbb{E}_{(x,y)\sim P_c}[\mathcal{L}_{\text{BCE}}(\mathcal{T}_t \circ F(x), y) + \gamma \mathcal{L}_{\text{CE}}(C_t \circ F(x), c)] \tag{5.3}$$

Then, the overall stream objective:

$$\mathcal{L}_{stream} = \mathcal{L}_{cluster} + \mathcal{L}_{target} \tag{5.4}$$

5.3.2 Memory Buffer Objective

We now present our approach to managing the memory buffer. During the execution of task $t \in \{1, 2, \ldots, T\}$, the buffer memory serves as a temporary storage area. Its capacity, denoted as $\mathcal{M}$, sets the maximum number of elements it can hold at any given time. Further, we can allocate a maximum number of elements for each task, known as the budget of the task $\mathcal{B}$. To select which example from the current task to insert, our insertion strategy considers the loss value – defined as $\mathcal{L}_{\text{BCE}}(x_i)$ – for the target label y. Afterwards, we use a set of intervals, called bins, to categorize these loss values into distinct ranges. An example within the memory buffer is hence associated with a specific bin, determined by the interval in which its loss value falls.

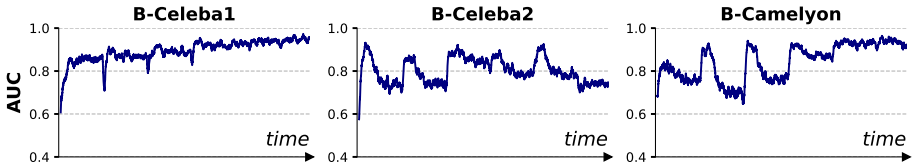


Figure 5.3: The AUC trend using binary cross-entropy loss to distinguish between the 'bias-aligned' and 'non-bias-aligned' groups. Notably, a higher AUC indicates that the loss is more effective at separating examples aligned with spurious correlations from those that are not. This result supports our strategy of achieving a balanced representation of bias-aligned and non-aligned groups, promoting fairer and more equitable sampling.

Buffer Management. As we initiate a task t, the budget allocation for the task is determined as $\mathcal{B} = \frac{\mathcal{M}}{t}$, taking considering the total capacity $\mathcal{M}$ and the task index t. The allocation for each bin is then defined as $\frac{\mathcal{B}}{n}$, which ensures a proportional distribution of memory resources across the predefined number of n bins. For each instance of data x_i in the training set D_t, a loss value $l_x = \mathcal{L}_{\text{BCE}}(\mathcal{T}_t \circ F(x), y)$ is calculated and stored, after a warm-up of a few epochs. This follows [129], which shows that the gap between bias-aligned and bias-not-aligned emerges during the initial training epochs. Thus, we choose to compute the loss after 5 epochs to leverage a larger gap. These values range from a minimum $\mathcal{L}_{\min}$ to a maximum $\mathcal{L}_{\max}$, which establishes the scope of the bins. The allocation of loss ranges to specific bins is determined based on their relative position within this range $\mathcal{L}_{\max} - \mathcal{L}_{\min}$, divided into n equal intervals.

Buffer Population. To determine whether to include an instance x in the buffer, we first check the current number of elements in its corresponding bin. If this number is below the allocated budget $\frac{\mathcal{B}}{N}$, the instance is included. This method ensures a fair representation of instances within the buffer, including both low and high loss values. This way, we ensure that the memory buffer always contains examples that are both aligned and not aligned with spurious correlations. Indeed, there is a significant empirical correlation between the loss value and potential biases (see Fig. 5.3). We leverage this correlation to maintain a balanced buffer.

Knowledge Distillation from Buffer Memory. Our implementation of knowledge distillation involves classifying samples stored in a buffer and comparing the cluster logits values saved to those computed for the current model. Let $y' = \mathcal{T}_t \circ F(x)$ represent the model output for task t (current or past), and $h' = C_t \circ F(x)$ represent the cluster classifier logits as well. We define two distinct terms:

$$\mathcal{L}_{\text{buf}} = \mathcal{L}_{\text{BCE}}(y', y) + \mathcal{L}_{\text{BCE}}(h', c), \tag{5.5}$$

The loss function $\mathcal{L}_{\text{buf}}$ combines the target and cluster classification loss. Addi-

Table 5.1: Average accuracy [↑] on B-CelebA1, B-CelebA2, and B-Camelyon. † BGS uses auxiliary group labels.

Method	B-CelebA1	B-CelebA2	B-Camelyon
Random	50.00± 0.00	50.00± 0.00	50.00± 0.00
SGD	60.12± 0.68	56.06± 0.09	85.80± 1.51
Debiasing			
BPA	61.69± 0.47	56.33± 0.53	88.06± 1.11
CFIX	64.00± 1.25	61.26± 0.96	87.88± 0.57
Replay (1024)			
BGS†	74.64± 0.34	76.45± 0.51	91.89± 0.41
ER-ACE	60.75± 0.77	59.03± 0.59	88.48± 0.08
DPP	61.34± 0.54	60.87± 0.36	82.56± 0.57
BPA + replay	60.92± 0.51	58.19± 0.66	88.88± 0.52
CFIX + replay	61.57± 0.39	62.62± 0.58	86.48± 0.58
LwP	62.33± 1.31	57.47± 1.47	86.39± 4.24
LwS (ours)	71.43± 1.67	70.73± 0.65	92.47± 0.21

tionally, we define the knowledge distillation objective for the buffer as:

$$\mathrm{KD}_{\mathrm{buf}} = \mathbb{E}_{(x,\bar{h})\sim\mathcal{M}}\left[\left\|\bar{h} - h'\right\|_2^2\right], \tag{5.6}$$

$\mathrm{KD}_{\mathrm{buf}}$ stands for the expected euclidean distance between stored logits $\bar{h}$ and the computed current logits h' over the distribution of samples $(x, \bar{h})$ drawn from the buffer memory $\mathcal{M}$. Finally, the overall objective function combines the stream, buffer, and knowledge distillation objectives:

$$\mathcal{L} = \mathcal{L}_{\mathrm{stream}} + \mathcal{L}_{\mathrm{buf}} + \mathrm{KD}_{\mathrm{buf}} \tag{5.7}$$

5.4 Experiments

Assessing debiasing methods in an environment affected by spurious correlations is challenging. Many works use synthetic datasets or custom splits to control a latent attribute z in a controlled setting [164, 168, 129, 206]. In our continual setting, we face a similar challenge as in [86, 102]. Here, we consider a sequence of tasks that occur successively, each influenced by a degree of bias. We extend the setting [102] by increasing the number of tasks.

To comply with standard metrics used in literature about debiasing [168, 164, 22, 206, 107, 155], we used the worst-case accuracy (not employed in [86]). Namely, we compute the average and worst accuracies across groups, where a group is defined as $g = (y, z)$. The group-specific accuracy is denoted as $acc_g(f_T, D_{test_t})$, representing the accuracy of the final model f_T on group g in

Table 5.2: Worst-group accuracy [↑] on B-CelebA1, B-CelebA2, and B-Camelyon. † BGS uses auxiliary (protected) group labels.

Method	B-CelebA1	B-CelebA2	B-Camelyon
Random	50.00± 0.00	50.00± 0.00	50.00± 0.00
SGD	14.87± 1.56	8.12± 0.57	48.53± 6.47
Debiasing			
BPA	15.08± 1.56	9.16± 0.47	62.13± 2.73
CFIX	18.00± 2.04	17.65± 1.97	59.56± 0.83
Replay (1024)			
BGS†	55.68± 2.92	56.56± 2.74	77.55± 0.07
ER-ACE	16.37± 1.76	13.12± 1.10	56.80± 1.70
DPP	21.79± 1.06	18.03± 1.37	53.40± 1.41
BPA + replay	16.04± 0.90	11.37± 0.43	65.33± 1.02
CFIX + replay	17.80± 0.04	19.79± 0.75	55.93± 0.34
LwP	19.40± 0.91	13.44± 3.94	54.40± 9.54
LwS (ours)	58.84± 2.42	50.91± 3.52	80.40± 1.74

the t-th task. The metrics for average and worst-group accuracies are defined as follows:

$$Acc_{avg}(f_T, D_{test}) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{|G|}\sum_{g\in G} acc_g(f_T, D_{test_t}) \tag{5.8}$$

$$Acc_{worst}(f_T, D_{test}) = \frac{1}{T}\sum_{t=1}^{T}\min_{g\in G} acc_g(f_T, D_{test_t}) \tag{5.9}$$

where G represents the set of all groups across tasks. Notably, each task t comes with its test unbiased dataset D_{test_t}, employed for evaluation.

Implementation Details. All reported results are the average of three runs. We use ResNet-18 [70] pre-trained on ImageNet-1K; **for fairness**, we apply this backbone to all tested methods. Each task was trained for 25 epochs using Stochastic Gradient Descent (SGD), with a learning rate of 1×10^{-3}. We performed k-means with $k = 8$ for all experiments. More details are provided in the supplementary.

5.4.1 Experimental Setup and Benchmarks

To model the presence of spurious correlations, we use CelebA [109] (facial attributes) and Camelyon17 [6] (skin lesions) from the WILDS benchmark [155]. We split the datasets into tasks such that a latent attribute z correlates with a target attribute, quantified by a given factor p_{corr}. We set p_{corr} to 0.95, indicating that 95% of images with a specific attribute y (*e.g.*, a necklace) are of a particular latent attribute z (*e.g.*, gender). In the supplementary materials, we provide an extensive graphical analysis of the correlation between the variables

y and z in our experimental settings. During training, we do not have access to the latent variables z; we use them only for evaluation.

Biased CelebA. The CelebA dataset [109] was divided into eight separate tasks for our study. These tasks focus on the binary classification of various target attributes y. We made two variants: **B-Celeba1** includes {*Heavy Makeup, Blond Hair, Receding Hairline, Young, Wearing Necklace, Bags Under Eyes, Smiling, Eyeglasses*} while **B-Celeba2** includes {*Chubby, Pale Skin, Bald, Gray Hair, Wearing Necktie, Wearing Hat, Arched Eyebrows, Mouth Slightly Open*}. Each task contains 4 480 images in the training set, evenly distributed in terms of y. The latent attribute z is the gender label as in [164, 22, 129]. Each task has a test data D_{test_t} with 100 samples per group (there are 4 groups for each task) to assess model debiasing performance.

Biased Camelyon. This dataset is derived from the Camelyon17 dataset [6]. It consists of 4 tasks, each involving binary classification. The hidden variable z represents the hospital from which the images were sourced. The presence of a tumor is indeed correlated with the hospital where the images were taken, thereby creating a spurious correlation between the two variables. The training includes 4 hospitals, while the test phase includes a fifth hospital not present in the training data. Each task contains 4,096 images, and the test sets are balanced for tumor presence and hospital origin, with 500 images per hospital.

5.4.2 Baseline and Competing Methods

Rehearsal Methods. While **SGD** does not incorporate measures against forgetting, **ER-ACE** [20] enhances traditional Experience Replay (ER) by applying distinct loss functions for the stream and the buffer. **DER++** [18] adopts self-distillation by encouraging consistency in the model's output, minimizing the L2 norm between the logits of current and past iterations. However, they do not consider the potential contamination of the buffer by spurious correlations, which could affect future knowledge retention.

Continual Debiasing Methods. To mitigate spurious correlations in both the stream and buffer, several methods have been proposed. **LwP** [86] aims to prevent spurious correlations by using self-supervised learning with feature-level augmentation. **BGS**† [102] constructs a buffer to store group-class-balanced examples across all encountered tasks. In this context, BGS acts as an oracle by leveraging latent-variable supervision z to structure the buffer.

Offline Debiasing Methods. We also assessed standard debiasing algorithms such as **BPA** [164], which employs a per-sample re-weighting strategy. **CFIX** [22] optimizes a dual objective to re-weight sample importance, using cluster classification as an additional regularization to smooth the latent space. Since these methods do not natively support the arrival of new tasks, we also introduce **BPA + replay** and **CFIX + replay**, which are adaptations that incorporate buffer reservoir sampling.

5.4.3 Experimental Results

Tab. 5.1 and Tab. Tab. 5.2 summarize the key findings of our work. LwS boosts average and worst-group accuracy metrics, outperforming rehearsal methods across various scenarios. A notable feature is the gain in worst-group accuracy, highlighting its effectiveness against spurious correlations. Also, the results demonstrate that our mechanism for updating the memory buffer enables the retention of unbiased past knowledge.

Baselines. Regarding debiasing methods, CFIX [22] and BPA [164] have effectively improved worst-case accuracy with respect to fine-tuning on the new task (SGD). However, their gains are relatively small compared to LwS, indicating the need for a buffer strategy to avoid forgetting. In this context, offline debiasing algorithms serve as more reliable baselines than naive fine-tuning (SGD).

Rehearsal Methods. Their results are reported in Tab. 5.1 and Tab. Tab. 5.2; we refer the reader to Fig. 5.4 for a in-depth comparison with DER++ [18], one of the most simple yet effective approaches. As shown, replay methods surpass their baselines, highlighting the advantage of memory replay. However, the tables reveal a crucial issue. If the buffer contains mostly biased elements, it can amplify bias in new tasks when samples are drawn from it. This underscores the limitation of traditional rehearsal methods.

Continual Debiasing Methods. From our results, LwS outperforms a continual debiasing model like LwP [86] and pairs the performance of BGS [102], which presents our upper bound. Indeed, it constructs the buffer **using supervision on the latent attribute** z to balance the number of elements per group in the memory, which is preferable but less realistic. Indeed, to identify the group labels, one must *i)* discover the variable z that determines the spurious correlation; *ii)* annotate the training set accordingly. This process is expensive and requires a thorough analysis of the dataset. Indeed, while annotating attributes like gender may be easy, it becomes impractical when the attribute z is hard to inspect (*e.g.* metadata protected by privacy laws or hidden artifacts in images). In such cases, a framework like ours, which avoids relying on group labels, is advantageous.

5.5 Ablation Studies

Reservoir Sampling Fails with Spurious Correlations. Tab. 5.3 illustrates the impact of memory buffer size ($\mathcal{M}$) and buffer handling strategies on LwS. The results reveal that the loss-based approach consistently outperforms the *reservoir* method in terms of worst-group and average accuracy across all datasets and buffer sizes (256, 512, 1024). This outcome supports our hypothesis that random strategies, such as *reservoir*, may unintentionally amplify spurious correlations
.

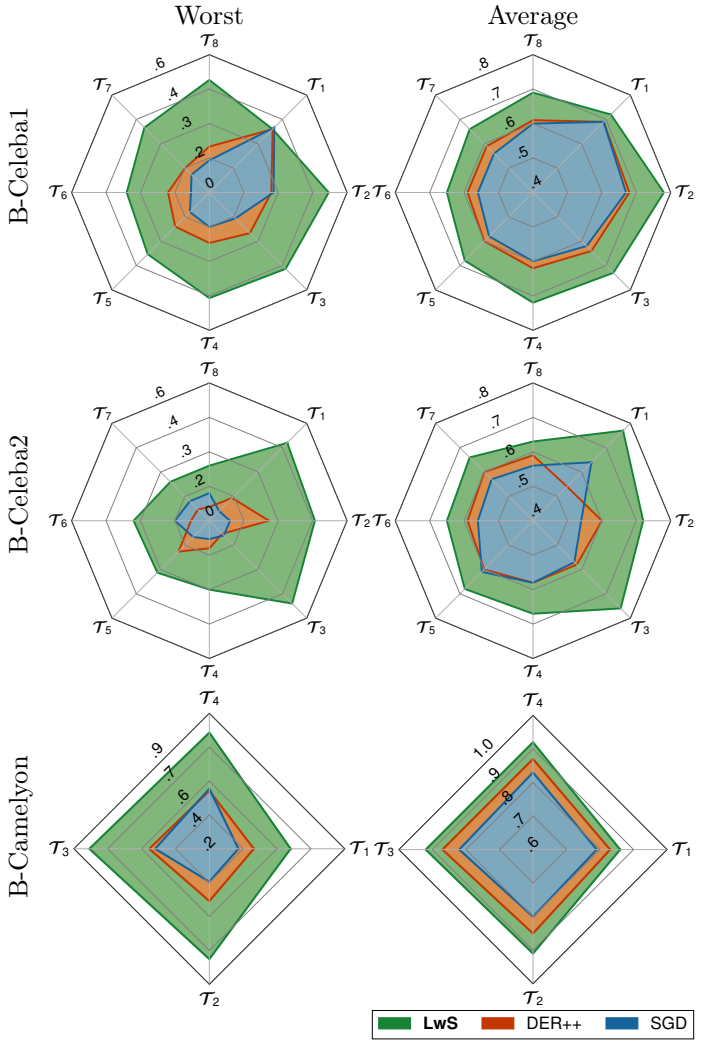


Figure 5.4: Comparative analysis across tasks, showcasing worst-group accuracy and average accuracy for each dataset.

Varying the Correlation Factor p_{corr}. We analyze how the model learns as the correlation factor changes and evaluate the effectiveness of different strategies. On the left side of Fig. 5.5, the relationship between the loss value and alignment with spurious signals (AUC) is shown as p_{corr} varies. After the warm-up phase, we compute the loss for all training elements of task t, which is then used in the buffer update described in Sec. 6.4. We observe a gradual decrease in AUC after

Table 5.3: LwS performance in terms of worst [↑] and average accuracy [↑] across different buffer sizes and management strategies.

	$\mathcal{M}$	Strategy	Acc_{worst}[%]	Acc_{avg}[%]
B-CelebA1	256	reservoir	14.14	58.21
		loss-based	**36.29**	**66.73**
	512	reservoir	18.50	61.08
		loss-based	**52.12**	**71.17**
	1024	reservoir	17.87	62.16
		loss-based	**56.98**	**72.57**
B-CelebA2	256	reservoir	18.71	61.10
		loss-based	**51.62**	**72.06**
	512	reservoir	19.37	62.43
		loss-based	**48.50**	**69.46**
	1024	reservoir	20.50	63.06
		loss-based	**53.37**	**71.40**
B-Camelyon	256	reservoir	41.40	81.92
		loss-based	**79.40**	**91.84**
	512	reservoir	36.80	81.92
		loss-based	**79.60**	**92.42**
	1024	reservoir	55.80	86.50
		loss-based	**80.40**	**92.84**

buffer insertion, as desired. As depicted on the right, joint training and DER++ are more susceptible to spurious correlations. As p_{corr} increases, both methods suffer a drop in average and worst-case accuracy while our approach performs robustly across different p_{corr} values.

On the Number of Bins. We investigate the effect of varying the number of bins for the buffer population. As the number of bins increases, we observe a slight decline in worst-case accuracy, as shown in Tab. 5.4. This trend can be attributed to the fixed buffer size; a greater number of bins entails a reduced allocation budget per bin, potentially leading to an under-representation of elements that diverge from the bias within each bin. Despite this, our strategy maintains competitive performance, even with a higher bin count.

Knowledge Distillation using Cluster Logits. We analyzed the impact of the KD_{buf} term introduced in Eq. 5.6. Our findings demonstrate that knowledge distillation offers significant advantages in smoothing the feature landscape and facilitating knowledge transfer across future tasks. In particular, Tab. 5.4

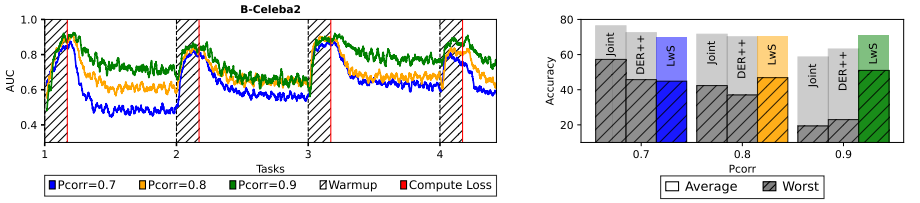


Figure 5.5: The figure displays AUC curves (left), which show the correlation between the loss value and alignment with spurious signals varying levels of p_{corr}. The shaded regions on the curves show the warm-up phase, followed by the computation of target losses for all training samples in each task. The right side presents comparisons across different p_{corr} values for joint training, DER++, and LwS methods.

Table 5.4: LwS performance comparison varying number of bins and usage of knowledge distillation (KD).

	B-CelebA1		**B-CelebA2**		**B-Camelyon**	
# bins	Acc_w	Acc_{avg}	Acc_w	Acc_{avg}	Acc_w	Acc_{avg}
2	58.23	72.71	55.12	75.40	76.40	90.88
4	**61.55**	**73.24**	53.37	71.40	81.20	92.72
8	53.12	70.79	51.25	70.68	**81.40**	92.40
16	55.61	71.83	52.00	71.72	80.40	92.84
32	50.37	70.82	51.00	71.34	79.60	**93.04**
no KD	58.80	**73.40**	50.50	70.18	80.40	**92.84**
w. KD	**61.55**	73.24	**53.37**	**71.40**	**81.20**	92.72

shows that utilizing cluster prediction logits improves the worst-case accuracy performance without negatively affecting the average accuracy.

On the Effect of w. Fixing $w_c = 1$ in Eq. 5.3 worsened model performance as shown in Tab. 5.5, demonstrating the effectiveness of our adaptive weighting strategy. As expected, the decrease with $w_c = 1$ was not severe thanks to the buffer population, which serves as a regularization term.

Sensitivity of γ. Tab. 5.6 shows how increasing the value of γ in Eq. 5.3 leads to better results. The scalar γ multiplies $L_{cluster}$ term, which indicates heterogeneity within a cluster c, where individuals share the same target label y (e.g., blond hair) but differ in attribute z (e.g., gender). Therefore, we assign a higher weight w_c to a cluster with a high expected error for L_{target} or $L_{cluster}$.

Table 5.5: LwS with adaptive weights w_c and fixed $w_c = 1$.

Dataset	w_c	Acc_{worst}	Acc_{avg}
B-CelebA1	adaptive	**58.84 ± 2.42**	**71.43 ± 1.67**
	fixed	52.83 ± 1.59	70.90 ± 0.99
B-CelebA2	adaptive	**50.91 ± 3.52**	70.73 ± 0.65
	fixed	47.29 ± 1.43	**70.74 ± 0.68**
B-Camelyon	adaptive	**80.40 ± 1.74**	**92.47 ± 0.21**
	fixed	78.20 ± 1.60	91.85 ± 0.37

Table 5.6: LwS results on B-CelebA1 using difference values of γ.

Metric	$\gamma = .0$	$\gamma = .2$	$\gamma = .5$	$\gamma = .8$	$\gamma = 1$
Acc_{worst}	47.61	51.92	55.62	54.85	**58.25**
Acc_{avg}	69.46	70.16	70.91	70.83	**72.12**

5.6 Discussion

Our results reveal a critical vulnerability in rehearsal-based continual learning: *the memory buffer can act as an amplifier of bias rather than a safeguard against forgetting.* When spurious correlations dominate the training distribution, random sampling strategies inadvertently populate the buffer with bias-aligned examples, creating a feedback loop that compounds the problem across tasks.

The success of LwS demonstrates that loss values serve as an effective proxy for detecting spurious correlations *without requiring explicit supervision.* As shown in Fig. 5.3, the separation between bias-aligned and non-aligned examples emerges naturally during early training epochs. By leveraging this phenomenon, we achieve performance comparable to BGS [102], which relies on expensive latent-variable annotations. Our ablation studies further confirm the limitations of reservoir sampling (Tab. 5.3), the robustness of LwS across varying p_{corr} values (Fig. 5.5), and the importance of cluster-based knowledge distillation for maintaining structural consistency (Tab. 5.4).

However, our approach has limitations. The reliance on k-means clustering assumes separable clusters in the feature space, (which may not hold for all domains), and the number of bins requires some tuning.

Overall, the work highlights a fundamental challenge in continual learning: mechanisms designed to preserve knowledge can inadvertently perpetuate biases. This has implications for deployed systems that learn continuously from biased data streams, where fairness may degrade over time.

6. Implicit Bias in Vision-Language Models

6.1 Overview

The debiasing approaches discussed in previous chapters primarily address explicit spurious correlations and observable disparities in model performance across predefined subpopulations, as measured by group-level accuracy metrics. However, vision-language models like CLIP [144], pre-trained on massive internet-scale datasets, inherit a more insidious form of bias: *implicit associations* that operate below the surface of standard fairness benchmarks. These biases manifest not through dramatic performance gaps, but through subtle stereotypical associations between visual concepts and social attributes.

Unlike explicit biases that violate measurable fairness constraints (*e.g.*, worst-group accuracy, demographic parity), implicit biases reflect *unintentional*, *uncontrollable*, and *purely stimulus-driven* associations [5, 8, 65]. Social psychology has long recognized this distinction: implicit attitudes differ fundamentally from explicit beliefs that individuals can consciously report. Drawing on this multidisciplinary perspective, we argue that the AI community requires evaluation frameworks grounded in psychological principles to capture biases that standard benchmarks overlook.

> *How can we measure implicit biases in vision-language models when they do not exhibit obvious performance disparities?*

Our investigation addresses this question by adapting established psychological measurement techniques to the domain of multimodal AI systems. Specifically, we leverage three frameworks from social cognition research: the *Stereotype Content Model* (SCM) [52], which organizes social perception along dimensions of Competence and Warmth; *Emotions Attribution* theory [68, 171], which examines affective responses toward social groups; and the *Semantic Differential Scale* [136], which captures connotative meaning through bipolar adjective pairs.

These instruments reveal hidden discrimination patterns—associations between demographic groups and evaluative attributes—that conventional accuracy-based metrics cannot detect.

To operationalize these psychological constructs for vision-language models, we introduce two complementary metrics. First, we adapt the *Common Language Effect Size* (CLES) [122], converting differential similarity scores into interpretable probabilities that quantify the extent to which models systematically associate certain attributes with specific demographic groups. Second, we develop a CLIP-adapted version of the *Implicit Association Test* (IAT) [65], replacing human reaction times with zero-shot classification preferences to measure the strength of stereotypical associations in the model's embedding space.

Using these metrics, we conduct a comprehensive audit of 90 Open-CLIP models [33] spanning diverse architectures (ResNet, Vision Transformers) and pre-training strategies (OpenAI, LAION, MetaCLIP, DataComp). Our analysis reveals that most models exhibit significant implicit biases along axes of ethnicity and gender, with systematic preferences for certain demographic groups when prompted with attributes related to Competence, Warmth, and emotional valence. Critically, we find that pre-training strategy profoundly influences the magnitude and direction of these biases, suggesting that dataset curation and training objectives play a decisive role in shaping implicit associations.

Beyond measurement, we investigate whether text-based debiasing techniques can mitigate these implicit biases. We evaluate orthogonal projection methods [35] that remove spurious directions from CLIP's text embedding space, comparing standard implementations against a novel variant that calibrates projections using SCM attributes. Our findings reveal a nuanced picture: while projection-based debiasing effectively reduces biases in already-biased models, it can paradoxically exacerbate disparities in models with initially weak biases, highlighting fundamental limitations of purely geometric interventions that ignore the semantic structure of social attributes.

6.2 Measuring Implicit Bias in CLIP

6.2.1 Common Language Effect Size (CLES)

We use the methodology developed for the Word Embedding Association Test (WEAT) [21] to evaluate bias in CLIP, which measures the differential association between two sets of target text concepts and visual embeddings. Here, A and B represent two sets of image embeddings of equal size (for example, white male and white female faces), and $x \in X$, a set of text embeddings which use a specific social attribute:

"A photo of a <adjective> looking face"

We define the cosine-similarity gap for a single text embedding x with respect

to sets A and B as follows:

$$\Delta_{gap}(x, A, B) = \left| \frac{1}{|A|} \sum_{a \in A} \cos(x, a) - \frac{1}{|B|} \sum_{b \in B} \cos(x, b) \right|, \tag{6.1}$$

Which is extended to a set of text embeddings X:

$$\Delta_{gap}(X, A, B) = \frac{1}{|X|} \sum_{x \in X} \Delta_{gap}(x, A, B). \tag{6.2}$$

This measure quantifies the differential association of the target concepts (text prompts) X with visual embeddings represented by A and B.

Interpreting Δ_{gap}: Effect Size as a Probability. Effect sizes are crucial in evaluating the outcomes of empirical studies. They determine whether an experimental intervention or manipulation yields a statistically significant effect and, if so, the magnitude of this effect. An example of effect size is Cohen's d, which is utilized to express the mean difference in terms of the standard deviations:

$$d = \frac{\mu_A - \mu_B}{\sqrt{\frac{(n_A-1)\sigma_A^2 + (n_B-1)\sigma_B^2}{n_A+n_B-2}}} \tag{6.3}$$

Cohen's d can theoretically range from 0 to infinity, with established benchmarks typically categorizing effect sizes as small ($d = 0.2$), medium ($d = 0.5$), and large ($d = 0.8$) [37]. However, these categories should not be rigidly applied as they are somewhat arbitrary, and even small effect sizes can be clinically significant in certain contexts [177]. An alternative measure is the Common Language Effect Size (CLES) [122], also known as the probability of superiority [66]. This statistic provides a more intuitive understanding than Cohen's d by converting the effect size into a percentage. It represents the probability that a randomly selected individual from one group will score higher than a counterpart from another group. There are two methods for calculating this probability: one is algebraic, while the other is empirical. The algebraic method assumes that the data is normally distributed and continuous, while the empirical approach does not rely on such assumptions [100].

Algebraic Approach. Mathematically, the CLES is the probability that a Z-score exceeds the value corresponding to no difference between groups in a normal distribution. Z-score can be calculated as follows:

$$Z = \frac{\Delta_{gap}(X, A, B)}{\sqrt{\frac{\sigma_A^2 + \sigma_B^2}{2}}}, \tag{6.4}$$

where $\Delta_{gap}(X, A, B)$ is the mean difference between the cosine similarities of groups A and B with respect to prompts X, and σ_A and σ_B are the standard

deviations of the cosine similarities within groups A and B respectively. The Z-score measures how the mean difference deviates from zero in terms of standard deviations.

The probability associated with this Z-score is calculated using the Cumulative Distribution Function (CDF) of the standard normal distribution. This gives the upper tail probability $P(Z > z)$, which represents the likelihood that $\Delta_{gap} > 0$:

$$P(Z > z) = 1 - \Phi(Z), \tag{6.5}$$

where $\Phi(Z)$ is the CDF of the standard normal distribution evaluated at Z. This probability quantifies the extent to which one group's embeddings are consistently rated as more similar to the prompts than the other's.

Empirical Approach. To avoid statistical assumptions, we measured the Common Language Effect Size (CLES) using the empirical method. This is accomplished by calculating the frequency with which $\cos(x, a) > \cos(x, b)$ holds true for all pairs (a, b) across all x in the set X.

6.2.2 Implicit Association Test (IAT)

Research in cognitive science [158] has led social psychologists to develop techniques for studying how individuals connect social groups with target concepts. A commonly used method is the Implicit Association Test (IAT) [65].

IAT with Humans. The IAT requires participants to quickly categorize items into different stimulus categories using one of two response keys. In an IAT focused on the racial attitudes of white individuals, four categories of stimuli might be used: pictures of black (ethnic out-group) and white (ethnic in-group) individuals, as well as positive and negative attributes. The IAT includes different experimental blocks: *(i)* a compatible block, where white individuals and positive attributes share the same response key, and black individuals and negative attributes share a different response key; *(ii)* an incompatible block, where these associations are reversed. The critical measure is reaction time —how long it takes to associate the pictures with the attributes. These experiments typically show that white participants are faster during the compatible block, associating white individuals with positive attributes and black individuals with negative attributes. This indicates a deep-seated in-group favoritism and out-group bias.

IAT with CLIP. We used a similar method to test the CLIP model, but reaction time was not a factor since it is constant. In the case of CLIP, the test involved zero-shot classification, using the similarity between the visual embeddings of the image and the textual embeddings of the input prompts. We used attributes from a semantic scale in our textual prompts to guide binary classification, such as positive versus negative.

Let $v(a) \in \mathbb{R}^d$ denote the normalized visual embedding of image a and $t(p) \in \mathbb{R}^d$ the normalized textual embedding of prompt p. Cosine similarity is computed as:

$$s(a,p) \;=\; \langle v(a),\, t(p) \rangle. \tag{6.6}$$

Let A and B denote the two image groups, and $\mathcal{P} = \{(p,n)\}$ the set of prompt pairs (*positive*, *negative*). The preference score for the positive prompt is defined as:

$$\mu_A \;=\; \frac{1}{|A|\,|\mathcal{P}|} \sum_{a \in A} \sum_{(p,n) \in \mathcal{P}} \mathbf{1}[\, s(a,p) > s(a,n) \,], \tag{6.7}$$

$$\mu_B \;=\; \frac{1}{|B|\,|\mathcal{P}|} \sum_{b \in B} \sum_{(p,n) \in \mathcal{P}} \mathbf{1}[\, s(b,p) > s(b,n) \,]. \tag{6.8}$$

The IAT score is computed as the absolute difference between the two rates:

$$\mathrm{IAT}_{\mathrm{score}} \;=\; \big|\, \mu_A - \mu_B \,\big|. \tag{6.9}$$

6.3 Debiasing CLIP from Text

Debiasing via Orthogonal Projection. It is essential for a robust classifier to avoid dependence on irrelevant features present in images. This necessitates the classifier to be invariant to image backgrounds or insensitive to attributes such as race or gender. To make the classifier invariant to irrelevant features, we utilize an orthogonal projection technique [35]. In such a scenario, matrix $M \in \mathbb{R}^{d \times m}$ represents the embeddings of spurious prompts, with the orthogonal projection matrix P_0 defined as:

$$P_0 = I - M(M^T M)^{-1} M^T, \tag{6.10}$$

where I is the identity matrix. Using P_0, we project text embeddings x to remove bias directions:

$$x_{new} = P_0 x. \tag{6.11}$$

Spurious prompts used to identify "bias" directions (matrix M) are:

"A photo of a male."	"A photo of a female."
"A photo of a man."	"A photo of a woman."
"A photo of a white person."	"A photo of a black person."

Figure 6.1: Representative samples from the dataset show individuals from different demographic groups.

Calibrating the Projection Matrix. Since P_0x could cause errors in estimating irrelevant feature directions, Chuang et al. [35] add a calibration term using a set of positive pairs of prompts S, which ideally retain the same semantic meanings post-projection. The calibration minimizes the following loss function, where λ is a regularization parameter:

$$\min_P \|P - P_0\|^2 + \frac{\lambda}{|S|} \sum_{(i,j)\in S} \|Px_i - Px_j\|^2, \tag{6.12}$$

resulting in the optimized projection matrix P^*:

$$P^* = P_0 \left(I + \frac{\lambda}{|S|} \sum_{(i,j)\in S} (x_i - x_j)(x_i - x_j)^T \right)^{-1}. \tag{6.13}$$

This process captures the pairwise differences $x_i - x_j$ for all pairs in S, refining the projection matrix to de-emphasize directions with larger singular values, enhancing the robustness of the debiasing process. Finally, the debiased embedding is then given by:

$$x_{new} = P^* x. \tag{6.14}$$

We refer to this method as *Orth Proj*.

Calibrating the Projection Matrix via Social Attributes. Building on the debiasing techniques detailed above, we further refine the calibration of the projection matrix, P^*, using the Stereotype Content Model (SCM) attributes discussed in Section 5.2. Typically, pairs of prompts in debiasing processes involve the same class of interest but include different spurious attributes. For example:

"A photo of a black male with dark hair." ≈ "A photo of a white male with dark hair."

In contrast, we define our class of interest using attributes from the SCM model, thereby aligning our debiasing efforts with sociopsychological insights. We refer to this method as *Our Orth Proj*. For instance, to calibrate P_0 as per Equation 6.12, we utilize prompt pairs such as:

"A photo of a competent looking black male." ≈ "A photo of a competent looking white male."

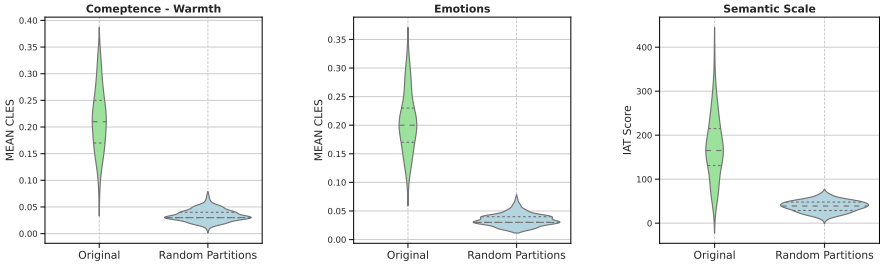


Figure 6.2: The plots show the distribution of the CLES and the IAT Score across models, comparing original and random data partitions.

6.4 Experimental Setup

6.4.1 Dataset

We used the Chicago Face Database (CFD) as a benchmark. The dataset includes males and females from various locations across the United States. Each person is shown with a neutral facial expression. For our experiments, we focused on 90 images per group (4 in total), all showing neutral facial expressions with closed mouths to minimize potential artifacts, as shown in Fig. 6.1.

6.4.2 Open-CLIP

Our experiments employed 90 Open-CLIP models [33]. The selected models include ResNet [70] and Vision Transformers (ViT) [45], such as RN50 and RN101, and various configurations of ViT (B-32, B-16, L-14, H-14). Other implementations, such as QuickGELU and specific model scales (e.g., ViT-B-32-256, ViT-H-14-378-quickgelu), were also explored. Selected models are pre-trained on distinct datasets and strategies including OpenAI [144], YFCC15M [176], CC12M [27], LAION [161, 160], Metaclip [195], DataComp-1B and CommonPool variations [59].

6.4.3 Evaluating Implicit Biases in Open-CLIP Models

Our analysis of the CLES and the IAT metrics across three distinct benchmarks of social psychology provides substantial empirical evidence against the null hypothesis. To model the letter, we conduct a permutation test using random equal-size partitions $\{(A_r, B_r)\}$ of $A \cup B$, which model the baseline assumption of no inherent bias in the associations between the groups and the visual-text inputs. In our experiments, rather than representing CLES as a value ranging from 0 to 1, we scale the metric to focus on the gap from the theoretical null hypothesis ($CLES = 0.5$), mapping it to $[0, 0.5]$. As depicted in Fig. 6.2, the

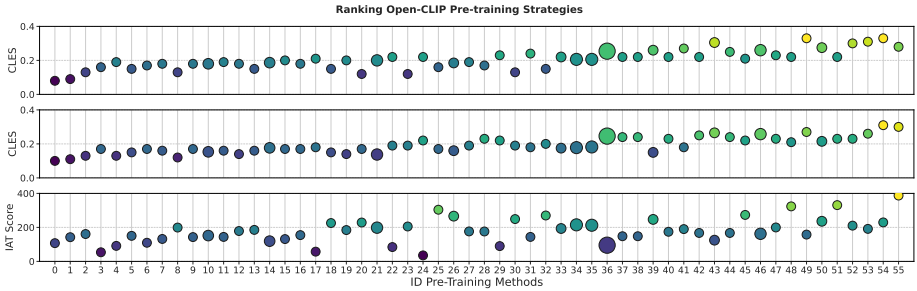


Figure 6.3: Ranking of Open-CLIP pre-training strategies based on bias metrics (SCM, Emotions, and Semantic Scale, respectively). Each dot represents a pre-training strategy and is color-coded to indicate relative performance, i.e., higher CLES values are closer to yellow. In comparison, lower CLES values are closer to the blue. The size of the dots is based on the number of models that use each strategy. The x-axis lists the IDs of the pre-training methods, ordered by the sum of ranks obtained across all metrics.

results indicate that the metrics obtained from the original data partitions significantly differ from those derived from random partitions. This gap confirms the presence of bias, which is consistently observed across 90 examined models.

On the Effect of Pre-training. We analyzed 56 different pretraining methods and found that each strategy had a distinct impact on social bias, as depicted in Fig. 6.3. The pretraining methods are ordered in the plot by their cumulative ranks across three metrics (SCM, Emotions, and Semantic Scale), providing a comprehensive view of how different training paradigms influence model behavior. This trend demonstrates the influence of pretraining method selection on the inclination toward discrimination, suggesting that the composition and curation of training datasets plays a decisive role in shaping the implicit associations learned by vision-language models.This observation underscores the importance of dataset quality over mere quantity and highlights the need for transparent documentation of data collection and filtering strategies.

Biased Image Retrieval. Considering the SCM attributes and using the worst and best-performing models, Fig. 6.4 plots the similarities between textual and visual embedding for all images. The plot reveals significant disparities, especially against images of Black individuals. Notably, except for the attribute "Warmth" where images of white women are most similar to the semantic meaning of the prompt, the model does not make significant distinctions at the attribute level. In this case, it shows a systematic preference for *White* individuals when prompted with attributes linked to *Competence* and *Warmth*, highlighting the need to address these biases in practical applications such as image retrieval.

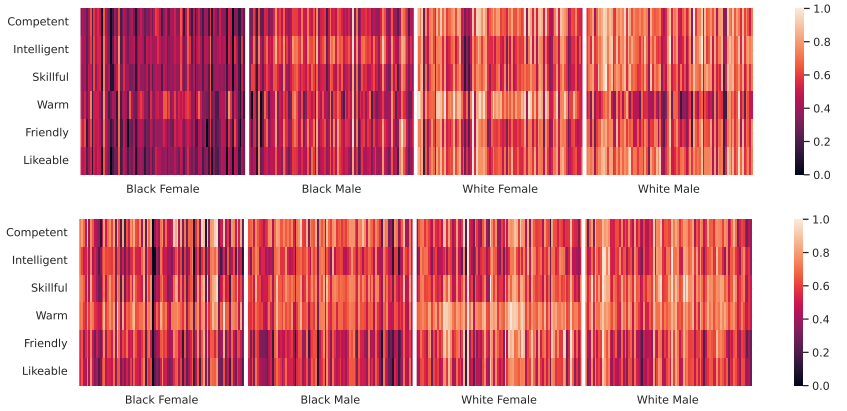


Figure 6.4: Visualization of biases in image retrieval tasks for different demographic groups. The heatmaps show the similarity score between the text prompt and the image for the worst-performing model (ViT-B-32 pre-trained by OpenAI, top) and the best-performing model (ViT-B-16 pre-trained with commonpool-l-text-s1b-b8k, bottom).

These findings have direct implications for downstream applications. In real-world image search systems, such biases could systematically surface images of certain demographic groups more prominently when users query for professional or leadership-related terms, thereby perpetuating and amplifying existing societal stereotypes.

6.4.4 Debiasing via Orthogonal Projection

Is Text-Guided Debiasing Enough? To assess the effectiveness of the debiasing strategies introduced in Sec. 6.3, Fig. 6.5 is provided. It shows that debiasing clip via orthogonal projection is primarily effective for models already exhibiting biased behavior. At the same time, it appears to saturate or even worsen the performance of less biased models. This paradoxical behavior suggests that purely geometric interventions, which operate by removing specific directions from the embedding space, may inadvertently disrupt the semantic structure that enables models with initially low bias to maintain balanced associations. The saturation effect observed in already-debiased models indicates that there exists a lower bound beyond which projection-based methods cannot improve fairness without sacrificing representational capacity.

Comparing Orth Proj with Our Strategy. Moreover, as expected, incorporating the attributes of the SCM model led to a systematic improvement. Our implementation improved the CLES in 47 out of 90 models, outperforming the *Orth Proj* [35], which improved in 33. Our approach enhanced performance in

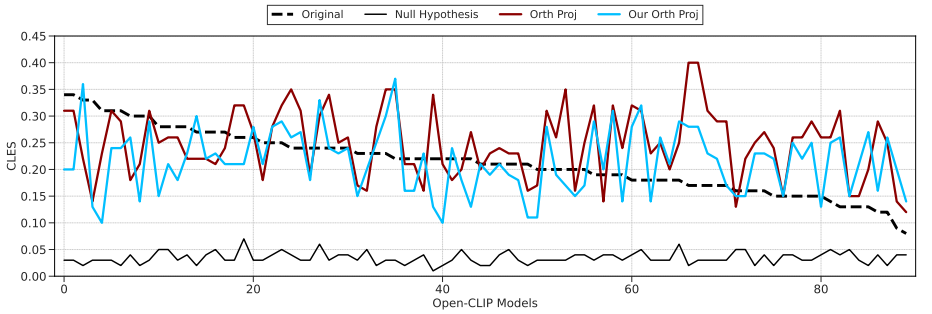


Figure 6.5: The CLES trends for debiasing strategies were analyzed across 90 Open-CLIP models. The results indicate that our implementation (cyan) improves upon *naive* one (red). However, while both debiasing strategies are mainly effective for models that exhibit bias, they tend to worsen the performance of less biased models.

64 out of 90 cases compared to the original *Orth Proj*, as shown in Fig. 6.5. The key difference lies in how we calibrate the projection matrix: by grounding the calibration process in psychologically validated social attributes from the SCM, we ensure that the debiasing procedure preserves semantically meaningful distinctions while removing spurious demographic associations. This demonstrates that incorporating domain knowledge from social psychology into the technical design of debiasing algorithms yields more robust and reliable outcomes than purely data-driven geometric approaches.

6.5 Discussion

In our study, we conducted a comprehensive analysis of implicit biases in open-clip models. Drawing from social psychology, we introduced two metrics: CLES and the adapted IAT. These metrics revealed significant disparities stemming from different visual inputs and demographics, highlighting the impact of visual data on skewing the embedding space and negatively affecting the alignment between text and image representations. We validated our results by adapting three different social psychology benchmarks to measure implicit bias in humans.

We found that the choice of pretraining significantly impacts such biases. We also evaluate debiasing methods that use orthogonal projection. Although these approaches have proven effective at reducing biases in models with apparent bias, they tend to exacerbate disparities in models with less obvious bias, highlighting their limitations. Overall, our study provides a new perspective on bias evaluation and emphasizes the ongoing need for scrutiny and refinement to ensure fairness and equity in such systems.

Book III

Aligning Parameter Spaces for Knowledge Transfer

7. Background

The previous chapters explored how neural networks fail under distribution shifts and proposed methods to mitigate shortcut learning in both static and continual learning scenarios. While continual learning aims to enable the integration of new information over time, it is only a partial solution to the broader challenge of model adaptability. Beyond mitigating the catastrophic forgetting, practical deployment requires addressing complementary questions: How easily can knowledge be transferred between models with different capacities or pre-training strategies? Can we compose capabilities from multiple specialized models without incurring the cost of retraining?

Recent discoveries reveal that seemingly distinct solutions found by independent training runs often reside in connected regions of the loss landscape [61, 55], and that algebraic operations on model weights can manipulate learned behaviors in semantically meaningful ways [82, 191]. These findings suggest that the weight space of neural networks exhibits exploitable symmetries, connectivity patterns, and compositional properties that can be leveraged to enable resource-efficient model updates. This chapter establishes the theoretical foundations for both **model alignment** via permutation symmetries and **model merging**, providing two complementary lenses for understanding and enabling parameter-space composition.

7.0.1 Mode Connectivity

Mode Connectivity occurs when paths of nearly constant loss connect different solutions within the loss landscape of neural networks (NNs) [61, 57, 61, 47]. When such paths are linear, we refer to *linear* mode connectivity (LMC) [55]. Entezari et al. [48] conjecture that solutions found by Stochastic Gradient Descent (SGD) can be linearly connected when accounting for permutation symmetries. Motivated by this, several works first align the models into a shared optimization space by permuting their neurons, and then merge them through a simple average [2, 90, 172, 140, 38, 130, 166], or apply optimal transport to align activations in transformer-based networks [83].

7.0.2 Weight Interpolation and Task Arithmetic

Emerging research reveals that the output of NNs can be manipulated through algebraic operations in weight space [80, 191]. Central to this paradigm are task vectors τ [82], which encode task-specific knowledge and exhibit compositional properties. More in detail, **task vectors** [82] are defined as the difference between fine-tuned and pre-trained weights: $\tau = \theta_{\text{fine-tuned}} - \theta_{\text{pre-trained}}$. These vectors capture task-specific knowledge in a form that exhibits remarkable compositional properties. Adding task vectors enables multi-task generalization, combining capabilities from multiple fine-tuning processes into a single model. Conversely, subtracting task vectors can selectively suppress learned behaviors, removing unwanted associations or reverting domain-specific adaptations, without degrading performance on unrelated tasks.

Beyond arithmetic, weight interpolations further unlock unexpected capabilities: blending fine-tuned and pre-trained weights often yields single-task performance superior to standalone fine-tuning [55, 85, 120, 148, 147, 192], suggesting a reconciliation of specialized adaptation with generalization capabilities. Multi-task merging via parameter averaging [80, 82, 191, 196] not only circumvents catastrophic forgetting [58, 121, 143] but synthesizes models that retain diverse expertise, even serving as superior starting points for future adaptation [34]. The benefits of weight ensembles and interpolations extend beyond just fine-tuned models; they also apply to models that are trained from scratch. Techniques such as those proposed by [2, 166], leverage permutation symmetries to facilitate coherent interpolation between models trained in different ways. Collectively, these findings position weight-space manipulation as a scalable toolkit for resource-efficient model engineering, where arithmetic and interpolation replace brute-force retraining.

7.1 Model Re-basin

Given two models with weights θ_A and θ_B, model re-basin [2] investigates how to permute the units of one model to facilitate the alignment of two models. The two models are then merged in the weight space, resulting in an interpolated model that achieves performance comparable to that of the two original ones.

Following the notation of [2], **re-basin** is defined as any operation defined on the weights of two models θ_A and θ_B that maps one of the two models onto the local loss region (*basin*) of the other one. To assess the effectiveness of re-basin, one common approach is to check the property of linear mode connectivity [55, 48] between the permuted model and the other, reference model. Informally, this involves checking if the model weights laying on the linear path connecting θ_A and θ_B also result in a low loss value.

To reach such a property, existing model re-basin techniques leverage the permutation symmetries inherent in neural networks [48]. These symmetries

allow the swapping of the units within a layer without changing the functionality of the network. To show that, we consider the activation of the ℓ-th layer of an MLP:

$$z_{\ell+1} = \sigma(W_\ell z_\ell + b_\ell), \quad z_0 = x, \tag{7.1}$$

where W_ℓ and b_ℓ are the weight matrix and bias vector and σ denotes an element-wise activation function. In this case, the following relation holds for any permutation matrix P:

$$z_{\ell+1} = P^\top P z_{\ell+1} = P^\top P \sigma(W_\ell z_\ell + b_\ell), \tag{7.2}$$

$$= P^\top \sigma(P W_\ell z_\ell + P b_\ell), \quad \text{where } P \in S_d, \tag{7.3}$$

with S_d denoting the set of $d \times d$ permutation matrices. Thanks to this relation, we can essentially permute the weights and biases of a layer using a matrix P. Therefore, when we apply the permutation P to the parameters of a layer, the resulting output undergoes the same permutation. However, to ensure that the transformed model remains functionally equivalent to the original, the next layer must process the output in its original, unpermuted form. This can be achieved equivalently by permuting the weights of the subsequent layer using the inverse permutation $P^\top$. Accordingly, we define a transformed set of weights θ' as:

$$W'_\ell = P W_\ell, \quad b'_\ell = P b_\ell, \quad W'_{\ell+1} = W_{\ell+1} P^\top. \tag{7.4}$$

Git Re-Basin. Ainsworth et al. [2] exploit Eq. 7.2 to induce weight alignment between θ_A and θ_B. Formally, we consider the ℓ-th feed-forward layer, with weight matrices $W_\ell^{(A)}$ and $W_\ell^{(B)}$ for θ_A and θ_B respectively. Given that each row of $W_\ell^{(A)}$ and $W_\ell^{(B)}$ represents a distinct feature, if $[W_\ell^{(A)}]_{i,:} \approx [W_\ell^{(B)}]_{j,:}$, then it makes sense to associate the units i and j. Therefore, we could formalize the alignment as finding the permutation matrix that maximizes the dot product between $P_\ell W_\ell^{(A)}$ and $W_\ell^{(B)}$. However, to preserve functional equivalence, we have to account for the term $P_{\ell-1}^\top$ related to the permutation of the previous layer —see Eq. 7.4. This results in a global optimization across layers:

$$\underset{\pi=\{P_\ell\}_1^L}{\operatorname{argmax}} \left\langle W_1^{(B)}, P_1 W_1^{(A)} \right\rangle + \left\langle W_2^{(B)}, P_2 W_2^{(A)} P_1^\top \right\rangle + \\ + \cdots + \left\langle W_L^{(B)}, W_L^{(A)} P_{L-1}^\top \right\rangle. \tag{7.5}$$

where $\langle A, B \rangle = \sum_{i,j} A_{i,j} B_{i,j}$ is the inner product between real-valued matrices. As discussed in [2], the optimization problem described in Eq. 7.5 corresponds to the Symmetric Orthogonal Bilinear Assignment Problem (SOBLAP), which is unfortunately NP-hard. Its relaxation re-casts it as a series of Linear Assignment Problems (LAPs), focusing on one permutation $\boldsymbol{P_\ell}$ at a time while keeping the others fixed. In formal terms:

$$\underset{P_\ell}{\operatorname{argmax}} \left\langle W_\ell^{(B)}, P_\ell W_\ell^{(A)} P_{\ell-1}^\top \right\rangle + \left\langle W_{\ell+1}^{(B)}, P_{\ell+1} W_{\ell+1}^{(A)} P_\ell{}^\top \right\rangle. \tag{7.6}$$

Notably, each LAP can be solved efficiently using polynomial-time methods such as the Hungarian algorithm [89]. The outcome is a set of permutation matrices $\pi = \{P_\ell\}_1^L$, which, when applied to model θ_A, result in a new model $\theta_{A'} = \pi(\theta_A)$. Notably, this model is functionally equivalent and, theoretically, lies within the low-loss basin of θ_B. However, since optimizing a series of LAPs is a coarse approximation of the SOBLAP, there are no strong guarantees of optimality.

In Chap. 8, we discuss model rebasin applied to transformers, showing that it is possible to transfer the knowledge of a task vector to a different backbone without additional data, using a permutation-based procedure.

7.2 Loss Landscape Properties

Weight-space operations such as model interpolation, re-basin transformations, and task vector arithmetic are post-training procedures that operate directly in a neural network's parameter space. A central question, therefore, is how the *local geometry* around a trained solution influences its ability to adapt in post-training updates. To do so, we introduce the concept of flatness and examine its role in measuring these properties.

7.2.1 Defining Flatness

Intuitively, the flatness of a minimum describes how sensitive the loss is around a solution θ^*. If the loss changes slowly, the minimum is considered *flat*; if it grows rapidly, the minimum is *sharp*. Flatness thus captures the local shape of the loss landscape in the neighborhood of θ^*. A classical way to formalize this notion is through the **Hessian** of second derivatives,

$$\mathbf{H} = \nabla^2 \mathcal{L}(\theta^*).$$

A solution is flat when the Hessian's largest eigenvalues are small, meaning the loss increases slowly in all directions or the ratio between the largest and smallest eigenvalues is low. However, computing or storing the full Hessian is intractable for modern neural networks. For large-scale models, one typically resorts to first-order curvature surrogates. A widely used proxy is the **Fisher Information Matrix** (FIM) [139], defined as:

$$\mathbf{F}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\nabla_\theta \log p(y \mid x;\theta)\, \nabla_\theta \log p(y \mid x;\theta)^\top\Big].$$

In practice, this expectation is approximated by an empirical average over dataset samples or mini-batches:

$$\widehat{\mathbf{F}} = \frac{1}{N}\sum_{i=1}^{N} \big(\nabla_\theta \log p(y_i \mid x_i;\theta)\big)\big(\nabla_\theta \log p(y_i \mid x_i;\theta)\big)^\top.$$

Usually, because estimating the full Fisher matrix requires estimating all pairwise correlations between parameters, one commonly adopts a *diagonal approximation*, which discards off-diagonal interactions and retains only the individual sensitivity of each parameter. This leads to the estimate:

$$\widehat{F}_{jj} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\partial}{\partial \theta_j} \log p(y_i \mid x_i; \theta) \right)^2 .$$

The diagonal Fisher therefore provides an efficient and informative proxy of local curvature: it preserves the primary signal needed to assess sharpness while reducing the computational and memory cost from quadratic to linear in the number of parameters.

7.2.2 Flatness and Robustness to Weight Perturbations

As discussed above, flat minima correspond to solutions that are robust to parameter noise: perturbing the weights within a wide, flat basin does not substantially change the network's behavior. This is why optimization methods such as SAM [54] and ASAM [99] explicitly promote flatness during training. A direct application of this concept is about quantization: indeed, we can interpret quantization as a perturbation of the weights, and the ability to be resilient to such operations can make the model usable even in settings with limited computational resources [128]. Similarly, in the context of **continual learning**, flat regions of the loss landscape provide learning stability: when the model lies within a wide basin, parameter updates due to new tasks are less likely to compromise performance on previous tasks [127, 125]. Consequently, the flatness of the loss landscape determines the extent to which the model can integrate new knowledge while preserving previously acquired capabilities.

7.2.3 Flatness and Model Merging

These observations naturally raise a critical question:

How do the geometry properties influence procedures like model-merging?

As demonstrated in Chap. 9, the geometric properties of the loss landscape play a crucial role in determining the success of downstream parameter-space composition. We show that independently fine-tuned models (task vectors) can be combined more reliably when their shared backbone converges to a flat basin around their respective optima. Flat regions naturally tolerate the weight perturbations arising from merging, increasing the likelihood that the composed solution remains functional. In contrast, sharp minima magnify even small discrepancies between models, causing the merged parameters to deviate rapidly toward high-loss regions and amplifying undesirable or unstable behaviors.

8. Re-Basin of Task Vectors

8.1 Overview

Recently, there has been a notable shift among researchers and practitioners towards fine-tuning pre-trained models, rather than building them from scratch. This method leverages backbones trained on large-scale datasets, significantly reducing the data and training time required to tailor models for specific downstream tasks. For this reason, since pre-trained backbones such as OpenAI's CLIP [144] are widely used as foundation models, their fine-tuned versions play a crucial role in numerous real-world applications, such as medical imaging [113] and satellite image analysis [117]. However, while these pre-trained backbones are widely adopted, their evolution poses new challenges, with tech companies and academic institutions frequently releasing updated checkpoints. Often, these updates do not modify the underlying architecture but involve new weights trained on increasingly large datasets compared to their predecessors [79]. Moreover, the additional training data may be more curated or specifically tailored to specialized domains, boosting their zero-shot capabilities considerably.

To take advantage of newly released checkpoints, the typical approach is to retrain them on the downstream task. This means fine-tuning the new checkpoint on the same data already used to adapt the original model. Besides the considerable costs associated with re-training the new model, this strategy is also unviable in certain scenarios. Indeed, the data for the downstream task might no longer be available due to compliance with privacy or storage constraints. This raises an important question:

> *Can we re-use the fine-tuning that has already been performed on the newly released model?*

Precisely, the overall aim of this paper is to investigate whether we can *transport* the previous fine-tuning, in a training-free manner. To understand the idea of *transport*, we consider the weights of the original base model as θ_A, and their fine-tuning as $\theta_A^{ft} = \theta_A + \tau$. The task vector [82, 135] $\tau = \theta_A^{ft} - \theta_A$ represents

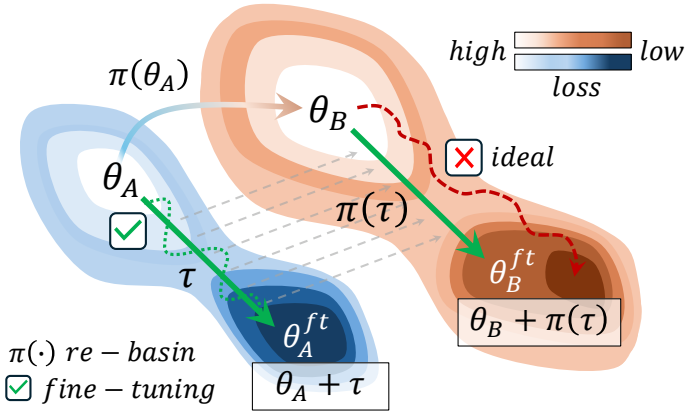


Figure 8.1: Transporting task vector τ from a fine-tuned base model $\theta_A^{ft} = \theta_A + \tau$ to a new release θ_B.

a direction from θ_A that embodies all adjustments made during the fine-tuning process. Hence, our goal is to find a procedure $\pi(\cdot)$ that can transport the task vector τ into an appealing basin of the newly released model θ_B (see Fig. 8.1). In this approach, the procedure $\pi(\cdot)$ must be designed to ensure that the modified weights $\theta_B^{ft} = \theta_B + \pi(\tau)$ achieve low loss on the downstream task.

When designing the transportation function $\pi(\cdot)$, the ideal approach should be data-free and training-free to meet the concerns above. Nevertheless, if the two base models θ_A and θ_B differ significantly (due to varying initialization, training strategies, or datasets), the knowledge acquired during fine-tuning of θ_A may not transfer to θ_B with a mere addition of the original task vector (*i.e.*, $\theta_B^{ft} = \theta_B + \tau$). To bridge the gap in representation spaces and facilitate the transfer, intuitively, we have to make the two base models "compatible", such that they "speak the same language". To address this challenge, we could *rebase* one of the two models (for instance, θ_A), such that any linear interpolation between the weights of the edited θ'_A and θ_B yields an intermediate model that performs comparably to both θ_A and θ_B. This indicates that the models are now aligned and thus share a common low-loss basin. Notably, this concept of re-basin models shares similarities with the approach described in [2], where alignment is achieved by finding optimal permutations of the rows in the weight matrices. We build upon this idea and explore its application in the context of fine-tuning, with a task vector being permuted and finally applied to θ_B.

While model re-basin presents an appealing framework, it currently faces several technical hindrances. To date, successful applications of model re-basin have been limited to Multi-Layer Perceptrons (MLPs) and Convolutional Neural Networks (CNNs) [2]. Unfortunately, the application of model re-basin to multi-

head attention layers, despite their widespread use in Transformer-based architectures, has been largely overlooked, with a few very recent attempts based on Optimal Transport (see Chap. 7). However, these recent methods do not guarantee **functional equivalence** between the permuted model and the original, unpermuted model. The primary obstacle lies in managing the weights associated with multiple attention heads. As we discuss in Sec. 8.2, indeed, to apply standard permutation-based approaches [2, 166, 83], the heads must be concatenated and treated as a single unified projection. This way, after applying permutations, each head of the edited model may incorporate rows from different original heads — an issue we refer to as *head contamination*. This is problematic because, without preserving the logical separation of heads during permutation, it becomes impossible to invert the permutation process and recover the original, unpermuted output of the attention block. Furthermore, existing methods struggle when two addends in the computational graph rely on distinct permutation matrices, a situation common in residual connections such as $h' = h + f(h)$. Differently, we avoid averaging the respective permutation matrices, thereby preserving their discrete nature.

To address these issues, we propose a structured two-level approach for effective re-basin of Transformer-based models, called TransFusion. To avoid *head contamination*, we first seek optimal mappings between pairs of heads (*inter-head* permutations); subsequently, we restrict permutations to only the rows within these coupled heads intra-head permutations. We mathematically prove that this two-level permutation strategy prevents head contamination and preserves **functional equivalence** between the original and permuted models. Notably, the inter-head permutations are optimized leveraging a distance metric that is invariant to permutations of the rows and columns within the heads. Such a metric is founded on spectral theory [91] and employs the singular values of the weight matrices, which are unaffected by orthogonal transformations like those induced by permutations. We show that transporting task vectors enables knowledge transfer to a new checkpoint in a data-free manner. In practice, this means we can improve the zero-shot performance of the new version on the downstream task. We also demonstrate that the transport retains generalization capabilities on a support set, a crucial factor for justifying updating the base model to the new release.

8.2 TransFusion: Re-basin of Task Vectors

Objective. Our approach, named TransFusion, is designed to transfer task-specific knowledge between transformer-based models that have undergone different pre-training. Specifically, it starts with an initial weight set θ_A and a task vector $\tau = \theta_A^{ft} - \theta_A$, derived after fine-tuning on a downstream task. The goal is to adapt τ to a new parameter configuration θ_B. This process aims to preserve the inherent properties of θ_B — for example, its superior zero-shot capabilities

compared to θ_A — and to integrate specialized knowledge carried out by τ for the downstream task. Finally, we aim to enable model transfer in a data-free manner without training. To achieve these objectives, we first align the weights of θ_A with those of θ_B. This is achieved with a novel *data-free weight-matching strategy* tailored for Transformer architectures. The procedure is discussed in depth in Sec. 8.2.1. Briefly, we address various shortcomings of existing methods and examine two building blocks of attention-based networks: residual paths and the multi-head attention mechanism. To manage the latter, we introduce a novel two-step process that employs a permutation-invariant spectral metric to match pairs of heads within the same layer of θ_A and θ_B. Subsequently, we permute features within the matched heads to optimize weight alignment, as detailed in Chap. 7.

Transport. We end up with a functionally equivalent model $\theta'_A = \pi(\theta_A)$, where $\pi(\cdot)$ yields a permutation of every layer in θ_A. Afterwards, $\pi(\cdot)$ is used to transport the task vector $\tau = \theta_{ft} - \theta_A$ into the low-loss basin of θ_B (see Sec. 8.2.2).

8.2.1 Attention Alignment for Transformer Models

A Transformer-based block consists of a multi-head attention layer and an MLP block, connected through residual connections. Considering the MLP, this builds upon standard linear projections, which we treat as discussed in Chap. 7. Instead, we adopt a novel, tailored approach for multi-head attention layers addressing a common pitfall. Considering multiple heads, current methods view their projections as a whole linear layer, thereby joining the corresponding weight matrices before applying permutations. However, such an approach does not reflect the organization of attention in distinct, parallel heads. For example, this can result in artifacts, where units from separate heads in the original model are mixed together — an issue we call *head contamination*. This compromises the structural separability of attention heads and precludes the preservation of *functional equivalence*, that is, the ability to permute and subsequently unpermute the weight matrices while yielding identical model outputs. In the following, we present our proposal against head contamination (see **1** and **2**) and a practical approach to handle residual connections (**3**). The complete methodology is outlined in Algorithm 1.

Step 1: Inter-Head Alignment. Consider the q(uery), k(ey), and v(alue) projection matrices W_q, W_k, and $W_v \in \mathbb{R}^{d_m \times d_m}$ — with d_m denoting the total embedding dimension of the attention module. We partition each matrix into $H = \#\text{heads}$ matrices (one for each head) of shape $d_k \times d_m$, where $d_k = \frac{d_m}{H}$. This results in a tensor $\tilde{W}_q = \text{split}(W_q, H) \in \mathbb{R}^{H \times d_k \times d_m}$ for the query projection matrix W_q. The same operation is applied for W_k and W_v to obtain $\tilde{W}_k$ and $\tilde{W}_v$.

The first step involves defining a distance metric between pairs of heads, such that we can identify and execute the optimal swap between heads in θ_A and θ_B

Algorithm 1 Weight Matching

Require: $\theta_A = \left\{W_\ell^{(A)}\right\}_{\ell=1}^{L}$ and $\theta_B = \left\{W_\ell^{(B)}\right\}_{\ell=1}^{L}$
Ensure: A permutation $\pi = \{P_1, \ldots, P_{L-1}\}$ of θ_A.

Initialize: $P_1 \leftarrow I, \ldots, P_{L-1} \leftarrow I$
repeat
  for $\ell \in 1, \ldots, L-1$ **do**
    if ℓ is a MHA layer **then**
      $P_\ell \leftarrow Algorithmmha_algo$
    else
      $P_\ell \leftarrow$ Solving LAP as in Eq. 7.6
until convergence

(see Fig. 8.2). We employ a distance metric that is **invariant** to permutations of rows and columns within the H sub-matrices in $\tilde{W}_q$, $\tilde{W}_k$, and $\tilde{W}_v$. In this respect, one might question why invariance is crucial for comparisons between different heads. We note that the initial, natural order of units does not necessarily correspond to the optimal alignment that could be achieved. Consequently, the metric used in this initial phase must be insensitive to the specific ordering of head features, thereby ensuring an agnostic comparison of the heads.

To achieve the permutation-invariance property, we employ a distance based on the **singular values** of the sub-matrices representing the heads. Specifically, given two heads $h_i^B = [\tilde{W}]_{i,:,:}^B \in \mathbb{R}^{d_k \times d_m}$ from model θ_B and $h_j^A = [\tilde{W}]_{j,:,:}^A$ from model θ_A, we compute the distance as:

$$d_{ij} = \|\Sigma_i - \Sigma_j\|, \tag{8.1}$$

where Σ_i and Σ_j denote the singular values of h_i^B and h_j^A respectively. These can be computed through the Singular Value Decomposition (SVD); in formal terms, considering the i-th head, the SVD decompose its weight $h_i^B = U_i \Sigma_i V_i^T$, where U_i and V_i are orthogonal matrices, and Σ_i is a diagonal matrix containing the singular values of h_i. As demonstrated in Sec. 8.5.1, the Euclidean distance between singular values remains invariant to permutations.

To take into account the distance for query, key and value projections jointly, we construct a distance matrix $D \in \mathbb{R}^{H \times H}$, where each element $D_{ij} = d_{ij}^q + d_{ij}^k + d_{ij}^v$ represents an inter-head alignment cost that is calculated as the sum of the pairwise distances across q, k and v matrices. We hence employ D to find the optimal inter-head permutation:

$$P_{\text{inter\_head}} = \underset{P \in S_H}{\operatorname{argmin}} \sum_{i=1}^{H} D_{i,P[i]}, \tag{8.2}$$

where $D_{i,P[i]}$ is the distance between the i-th head of model θ_B and the $P[i]$-

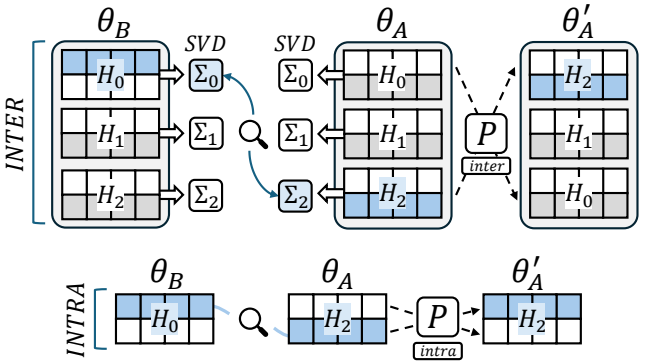


Figure 8.2: Inter- (Step 1) and intra-head alignment (Step 2).

th (candidate) head of model θ_A. The solution $P_{\text{inter\_head}}$ can be practically determined with the Hungarian algorithm. The corresponding permutation is applied to each $\tilde{W}_q^A$, $\tilde{W}_k^A$, and $\tilde{W}_v^A$, thereby reordering the heads of θ_A to increase alignment with those of θ_B.

Step 2: Intra-Head Alignment

After matching each pair of heads $(h_i^B, h_{P[i]}^A)$, we aim to swap individual units between h_i^B and $h_{P[i]}^A$. To do so, as in Git Re-Basin [2], we seek for permutations that maximize the inner product between the corresponding H sub-portions of the projection weights h_i^B and $h_{P[i]}^A$, as follows:

$$P_{\text{intra\_head}} = \{P_{\text{intra\_head}}^{(i)}\}_{i=1}^{H},$$
$$\text{where } P_{\text{intra\_head}}^{(i)} = \underset{P \in S_{d_k}}{\operatorname{argmax}} \left\langle h_i^B, P h_{P[i]}^A \right\rangle. \tag{8.3}$$

In this formula, the cost is computed as the dot product across the query, key, and value sub-matrices. In summary, this two-step process ensures both global inter-head and local intra-head alignment while preserving the structural integrity of the multi-head attention mechanism. The comprehensive procedure for aligning multi-head attention layers, combining these alignment stages, is formalized in Algorithm 2. These individual alignment steps are unified into a single composed permutation, denoted as P_{attn}, which is applied directly to the projection matrices of the multi-head attention layer. Crucially, it can be shown that our structured composed permutation preserves functional equivalence: despite reordering and permuting the heads and their internal dimensions, the attention computation remains unchanged. We formalize this in the following theorem.

Theorem 8.2.1 (Equivariance of Multi-Head Attention) *Let P_{inter_head} be a permutation over the H attention heads, and let P_{intra_head} be a set of*

Algorithm 2 Attention Alignment

Require: Weights $\tilde{W}_q^{(A)} P_{\ell-1}^\top$, $\tilde{W}_k^{(A)} P_{\ell-1}^\top$, $\tilde{W}_v^{(A)} P_{\ell-1}^\top \in \theta_A$ and $\tilde{W}_q^{(B)}$, $\tilde{W}_k^{(B)}$, $\tilde{W}_v^{(B)} \in \theta_B$ for multi-head attention projection layer ℓ and previous layer $\ell - 1$.

Ensure: Permutation $P_{\ell\text{-attn}}$ for $\tilde{W}_q^{(A)}$, $\tilde{W}_k^{(A)}$, $\tilde{W}_v^{(A)}$.

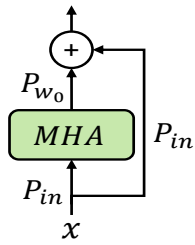
Step 1: Inter-Head Alignment
Create spectral distance matrix D (Eq. 8.1).
$P_{\text{inter}} \leftarrow$ Solve LAP on D (Eq. 8.2).
Step 2: Intra-Head Alignment
for $h = 1$ **to** H head pairs from P_{inter} **do**
  $P_{\text{intra}}^{(h)} \leftarrow$ Solve LAP for head pair h (Eq. 8.3).
$P_{\ell\text{-attn}} \leftarrow P_{\text{inter}} \circ \left\{P_{\text{intra}}^{(h)}\right\}_{h=1}^{H}$ ▷ compose permutations

independent permutations acting within each head (of size $d_k = \frac{d_m}{H}$). Then applying the composed block permutation P_{attn} to each of the projection matrices $W_q, W_k, W_v \in \mathbb{R}^{d_m \times d_m}$ is functionally equivalent to permuting the output of the multi-head attention module. The resulting attention output O' satisfies: $O' = OP_{attn}$, where O is the unpermuted output. (A complete proof of Theorem 8.2.1 is provided in Sec. 8.5.2.)

Step 3: Managing of Residual Connections

Each transformer block incorporates two residual connections: the first bypasses the multi-head attention layer, and the second bypasses the feed-forward network:

$$\begin{aligned} \mathbf{z}_{attn} &= W_0\,\mathrm{MHA}(\mathbf{x}), \\ \mathbf{z}_i &= \mathbf{z}_{attn} + \mathbf{x}, \\ \mathbf{z}_f &= W_2 \mathrm{ReLU}(W_1 \mathbf{z}_i), \\ \mathbf{z}_{out} &= \mathbf{z}_f + \mathbf{z}_i, \end{aligned} \tag{8.4}$$



where $\mathbf{x}$ is the input, W_0 is the weight of the attention mechanism, and W_1 and W_2 those of the feed-forward layer. For simplicity, we omit layer normalization as it can be regarded as a standard linear projection. In each residual block, the input and the intermediate output are summed to produce the final output. However, we note that there are several sources of potential mismatch between the two addends: intuitively, if the two addends have undergone different permutations, it is reasonable to suspect a potential mismatch in their representations.

To clarify the interaction between permutations in residual blocks, consider Sec. 8.2.1, which represents the first residual $\mathbf{z}_i = \mathbf{z}_{attn} + \mathbf{x}$. When the

weights of the model are permuted, the input $\mathbf{x}$ comes with its own permutation P_{in}, which has to be accounted for using $P_{\text{in}}^{\top}$. Moreover, the attention projection W_0 adds its own permutation matrix P_{W_0}. This leads to the following relation:

$$\mathbf{z}_i = P_{W_0}\mathbf{z}_{\text{attn}} + P_{\text{in}}\mathbf{x}. \tag{8.5}$$

When examining the permutations/summations that impact on $\mathbf{z}_i$, there are two main issues: *issue I)* the residual branch lacks transformations that could account for the matrix P_{in}; *issue II)* the projection W_0 adds its own permutation P_{W_0} of which the residual branch has no information about.

To maintain coherence between the two addends, they must be transformed under identical permutations. To enforce this consistency, we redefine the identity mapping made by the residual connection. We replace it with a composition, $\mathcal{I}_i = P_{W_0}P_{\text{in}}^{\top}$, consisting of two permutations — one to address *issue I* and another for *issue II* — as follows:

$$\mathbf{z}_i = P_{W_0}\mathbf{z}_{\text{attn}} + \mathcal{I}_i P_{\text{in}}\mathbf{x} = P_{W_0}\mathbf{z}_{\text{attn}} + P_{W_0}\mathbf{x}, \tag{8.6}$$

which highlights how the two addends now share the same permutation. An analogous process applies to the second residual connection yielding $\mathbf{z}_{\text{out}}$ (see Sec. 8.5.3 for the full procedure). As a final technical note, we remark that the permutation matrix P_{W_2} associated to the second residual block in Eq. 8.4 has to be considered as input permutation for the subsequent layer.

8.2.2 Transporting Task Vectors from θ_A to θ_B

By applying π to model θ_A, we would have a functionally equivalent model $\theta'_A = \pi(\theta_A)$ with stronger linear-mode connectivity with θ_B compared to the original θ_A. However, to allow knowledge transfer from the fine-tuned model $\theta_A^{ft} = \theta_A + \tau$ to θ_B, we do not apply the permutations directly on θ_A, but rather on the task vector τ, as follows:

$$\text{task vector}: \quad \tau = \theta_A^{ft} - \theta_A, \tag{8.7}$$

$$\text{transport}: \quad \tilde{\theta}_B^{ft} = \theta_B + \alpha\pi(\tau), \tag{8.8}$$

where α is a non-negative scaling factor [192] modulating the influence of $\pi(\tau)$ on θ_B.

By leveraging the concept of transporting task vectors, we have several notable advantages, especially in a scenario with multiple models fine-tuned on distinct tasks from the same base model θ_A. In this scenario, the weight matching process between θ_A and θ_B needs to be conducted only **once**. Indeed, a permutation set π can be established and reused to transfer any number of task vectors. This approach avoids the additional computational costs associated with learning separate transport functions for each transfer. Moreover, transporting multiple task vectors using the same reference model θ_A allows their combination at destination θ_B, which basically means we could still apply model merging [191] after re-basin.

8.2.3 Complexity Analysis

In this subsection, we assess the computational complexity of the proposed weight matching procedure. The key insight is that the method is highly efficient compared to full re-training, and scales polynomially with model size.

Proposition 8.2.2 *Let L be the number of layers and d_m the embedding dimension of each transformer block. The overall computational complexity of our weight matching procedure is dominated by $O(L\, d_m^3)$. This complexity matches that of Git Re-Basin, making our approach comparably efficient in terms of computational cost.*

The proof is provided in Sec. 8.5.4 and illustrates the per-layer contribution of both MLP and attention blocks.

8.3 Experiments

This section is structured into three main parts. Initially, we empirically assess the transportation of task vectors, involving extensive experiments across both visual and natural language processing (NLP) tasks (Sec. 8.3.1). Subsequently, we examine the capability of our methodology to align the weights of two Transformer models while maintaining functional equivalence (Sec. 8.3.2). Finally, several ablative studies show the impact of our techniques on addressing multi-head attention layers and residual connections (Sec. 8.3.3).

8.3.1 TransFusion of Task Vectors

Visual Classification Tasks. As reference architecture, we consider the CLIP ViT-B/16 Vision Transformer [144] from Open-CLIP [33]. We refer to θ_A as the original pre-training weights and θ_B as those used for the re-basin. We use CommonPool pre-training for θ_A and Datacomp for θ_B, both cited in [59].

Considering the base model θ_A, we fine-tune the corresponding model on several computer vision tasks [144, 82]. We employ DTD [36], EuroSAT [71], GTSRB [170], and SVHN [131] and obtain multiple, independent fine-tuned models like $\theta_A^{ft} = \theta_A + \tau$. Afterwards, we empirically assess the transportation of τ to the new weights θ_B. In this respect, we adopt two metrics to characterize the quality of the transported model $\theta_B + \pi(\tau)$: *i)* the zero-shot performance on the original task (*specialized knowledge*), and *ii)* the zero-shot performance on a support, unseen set to evaluate the preservation of *broader capabilities*. In our experiments, ImageNet-R [72] serves as a support dataset.

We report the results in Tab. 8.1 as drops (-) or gains (+) in accuracy relative to the zero-shot performance of θ_B. As baselines, we provide the results of *vanilla transportation* (no permutations applied on τ) and those of Git Re-Basin [2] and

Table 8.1: Comparison of permutation-based methods on visual tasks, in terms of task accuracy [↑] and support accuracy [↑].

Method	**EuroSAT**		**DTD**		**GTSRB**		**SVHN**	
	Task	Supp.	Task	Supp.	Task	Supp.	Task	Supp.
θ_B *zero-shot*	49.02	68.73	47.50	68.73	43.42	68.73	45.97	68.73
$\theta_B + \tau$	-7.62	-16.15	-0.15	-0.10	-5.39	-0.70	-22.00	-16.45
$\theta_B + \pi\tau$*(OT)*	-14.05	-5.28	-0.53	-1.18	-2.43	-1.30	-12.30	-2.70
$\theta_B + \pi\tau$*(Rebasin)*	+0.95	-0.48	-0.91	**-0.02**	+0.76	**-0.05**	+0.79	**+0.30**
Ours	**+4.95**	**-0.06**	**+0.21**	-0.08	**+1.10**	-0.40	**+3.64**	-0.48

Table 8.2: Comparison of permutation-based methods on NLP tasks, in terms of task accuracy [↑].

Method	**QQP**	**SST2**	**RTE**	**CoLA**
θ_B	55.00	50.69	54.51	40.94
$\theta_B + \tau$	-8.29	+0.23	-2.53	-0.77
$\theta_B + \tau$ *(OT)*	-8.31	+5.39	-1.08	-1.25
$\theta_B + \tau$ *(GiT Re-Basin)*	+3.58	+5.73	+2.17	+1.44
TransFusion (Ours)	**+6.50**	**+5.96**	**+3.61**	**+2.49**

Optimal Transport (OT) [83], two existing methods for model re-basin. Specifically, the comparison with OT is noteworthy since this approach is designed for Transformer models (like ours).

As can be seen, our method enhances zero-shot performance on the downstream tasks and preserves generalization on the support dataset, outperforming existing permutation-based methods. Considering the results of our approach, it is particularly noteworthy that we enhance performance on the downstream task while maintaining generalization, all achieved without the use of any data.

In the experiments shown in Tab. 8.1, we consistently set the scaling coefficient for the (permuted) task vector as $\alpha = 1$ (see Sec. 8.2.2). This illustrates the drop/gain in accuracy for $\theta_B + \alpha\tau$ (blue) and our $\theta_B + \alpha\pi(\tau)$ (red). This drop/gain is measured w.r.t. the zero-shot accuracy of θ_B, and α varies within the range $[0.01, 2.0]$. The outcome is that applying the permuted $\pi(\tau)$ to θ_B leads to tangible improvements in the downstream task (top row), especially $\alpha \approx 1$. Moreover, when $\alpha \geq 0.5$, the permuted task vector is considerably more reliable in terms of generalization (higher accuracy on the support set).

NLP Classification Tasks. Herein, we investigate a different setting that involves closed-vocabulary text classification — specifically, a set of tasks from the GLUE benchmark [187]. We consider a model $\theta = \{\phi, \omega\}$ composed of a pre-

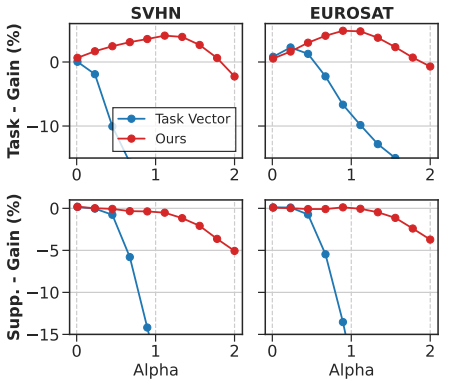


Figure 8.3: Zero-shot gain/drop relative to θ_B of naive $\theta_B + \alpha\tau$ (blue) and our strategy $\theta_B + \alpha\pi(\tau)$ (red) varying α.

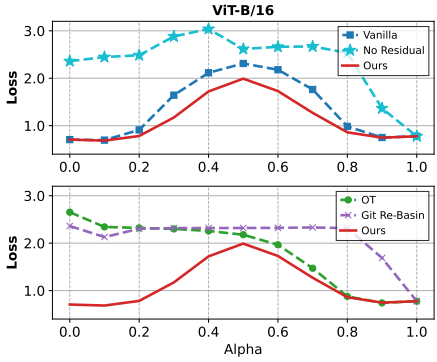


Figure 8.4: Loss barrier on CIFAR-10. Top: Our permutation approach vs. vanilla interpolation and no residual variant. Bottom: Comparison with methods, which fail to preserve functional equivalence as $\alpha \to 0$.

trained Transformer encoder ϕ and a classification head ω. We then evaluate the transport of the learned task vector $\tau_\phi = \phi_A^{ft} - \phi_A$ on a new feature extractor ϕ_B. As access to data of the downstream task is restricted, we are unable to train a new classifier for θ_B: consequently, we re-use the originally fine-tuned classifier, denoted as ω^{ft}. The goal is to evaluate whether transporting the task vector τ_ϕ aligns the representation yielded by $\phi_B + \pi(\tau_\phi)$ with the original, fine-tuned classifier ω^{ft}.

Tab. 8.2 presents the evaluation for the GLUE benchmark. Unexpectedly, applying the classification head from the original feature extractor ϕ_A yields poor performance (see first line of Tab. 8.2, θ_B). On the other hand, transporting τ_ϕ with Git Re-Basin and Optimal Transport performs reasonably, with good gains on QQP and SST2. Moreover, our approach leads to the highest and more consistent performance gains, highlighting the potential of our framework.

8.3.2 TransFusion Improves Alignment and Preserves Functional Equivalence

While the previous analyses focus on transferring task vectors, we now delve into the effectiveness of our approach in terms of weight alignment. In detail, we consider two ViT-B/16 models [45] A and B trained independently on CIFAR-10 [98] from scratch, which means they underwent different initializations and batch orders. After training, we apply our permutation strategy to θ_A and analyze the resulting alignment of $\pi(\theta_A)$ and θ_B in terms of *linear mode connectivity*:

following [2], we evaluate the loss landscape $\mathcal{L}\left((1-\alpha)\pi\left(\theta_A\right)+\alpha\theta_B\right), \alpha \in [0,1]$ while interpolating between the two models.
As shown by Fig. 8.4 (top), applying the permutation π yields an interpolated model that exhibits consistent lower loss compared to the vanilla approach that does not permute θ_A ($\pi(\theta_A) = \theta_A$). Moreover, the *no residual* approach underscores the critical role of properly handling residual connections in our method — both during the permutation of θ_A and throughout the interpolation process between models. Similarly, in Fig. 8.4 (bottom) we assess the loss landscape using π derived from the Optimal Transport (OT) hard alignment method [83] and Git Re-Basin [2]. While these works show a more favorable loss landscape than naive interpolation, we observe that the resulting interpolated model struggles and exhibits high loss as $\alpha \to 0$, highlighting that neither OT nor Git Re-Basin preserves functional equivalence. We conjecture that this lack stems from potential weaknesses in effectively permuting layers that feature residual connections and multi-head attention blocks. In light of the results achieved by our approach, we claim that it represents the first successful data-free method to interpolate between two Transformer models in weight space, while ensuring the functional equivalence of $\pi(\theta_A)$.

8.3.3 Ablative Analysis

On the Strategy to Manage Multiple Heads. We herein explore the significance of an appropriate policy for permuting the projection layers within multi-head attention mechanisms. Specifically, we present the outcomes of transferring τ with varying strategies to permute attention projection layers. As potential alternatives, we firstly consider **brute force alignment**, which considers all possible head pair combinations within each attention layer. Then, for each candidate pair, the *intra-head* alignment cost is computed by optimizing the objective in Eq. 8.3. The final permutation is then derived with the Hungarian algorithm, which selects pairs with the highest intra-alignment scores.

After, we compare our approach with one that pairs heads in A and B according to their natural order, thereby avoiding *head contamination* by design. Nevertheless, the preservation of original head ordering comes at the cost of ignoring functional mismatches between attention units. We refer to this further baseline as **no attention alignment**.

The results of these experiments are detailed in Tab. 8.4. Our attention-alignment strategy achieves superior performance on the downstream task ($\theta_B + \pi(\tau)$) compared to alternative approaches, while maintaining comparable zero-shot capabilities. The comparison with the brute force approach underscores the effectiveness of our permutation-invariant costs in modeling inter-head relationships, demonstrating superior performance over a brute force alignment that optimizes for the best match within each candidate pair of heads. Furthermore, our results suggest that preserving the original head ordering (as in

Table 8.3: Fine-tuning results with permuted task vectors.

METHOD	EUROSAT	DTD	GTSRB	SVHN
θ_B ZERO-SHOT	49.02	47.50	43.42	45.97
$\theta_B + \alpha\tau$	+7.93	-1.44	+4.70	-15.98
$\theta_B + \alpha\pi(\tau)$	**+10.00**	**+1.21**	**+6.80**	**+10.52**

Table 8.4: TransFusion results by head alignment strategy.

HEAD ALIGN.	EUROSAT		GTSRB		SVHN	
	TASK	SUPP.	TASK	SUPP.	TASK	SUPP.
θ_B ZERO-SHOT	49.02	68.73	43.42	68.73	45.97	68.73
BRUTE FORCE	+1.32	-0.21	+0.60	-0.46	+3.39	-0.40
NO ATT-ALIGN	+2.22	-0.47	+0.71	**+0.05**	+0.24	**-0.08**
OURS (FULL)	**+4.95**	**-0.06**	**+1.10**	-0.40	**+3.64**	-0.48

the no-alignment strategy) yields better performance than brute-force inter-head matching for two of the three experimental tasks.

Few Shot Fine-tuning. There are practical scenarios in which retaining data is infeasible. If such constraints are not present, our proposed method can be effectively combined with fine-tuning. To illustrate this, we follow [205] and start with a small subset consisting of 10 shots per class, learning a scaling coefficient per layer, denoted as $\alpha = [\alpha_1, \ldots, \alpha_{|L|}]$. The results in Tab. 8.3 clearly indicate a substantial improvement when fine-tuning a model that has undergone re-basin using our approach, represented as $\theta_B + \alpha\pi(\tau)$. In contrast, fine-tuning directly from $\theta_B + \alpha\tau$, without permutation, yields inferior outcomes. This emphasizes that re-basing and fine-tuning should not be considered mutually exclusive but complementary strategies.

8.4 Discussion

As pre-trained checkpoints are frequently updated, the ability to transport fine-tuning in a data-free manner becomes valuable, particularly when data retention is constrained by privacy or storage limitations.

The success of TransFusion demonstrates that weight-space alignment is a viable alternative to reusing fine-tuned Transformers. Our two-level permutation strategy, combining spectral-based inter-head matching with intra-head alignment, preserves functional equivalence during re-basin as evidenced by the loss barrier in Fig. 8.4). The method's effectiveness depends on the quality of source fine-tuning: strong task vectors (GTSRB, EuroSAT) yield substantial gains, while weaker ones (DTD) provide modest improvements.

8.5 Proofs

8.5.1 On the Invariance to Permutations of our Metric for Inter-head Alignment

Proposition 8.5.1 *Let $h \in \mathbb{R}^{m\times n}$ be arbitrary. For any h, denote its singular values by $\sigma(h) = (\sigma_1(h), \sigma_2(h), \ldots, \sigma_{\min(m,n)}(h))$, where $\sigma_1(h) \geq \sigma_2(h) \geq \cdots \geq 0$. For two matrices h_1, h_2 of the same shape, define*

$$d_p(h_1, h_2) = \|\sigma(h_1) - \sigma(h_2)\|_p \,, \tag{8.9}$$

where $\|\cdot\|_p$ is the usual p-norm for vectors. Then, for any permutation matrices $P_r \in \mathbb{R}^{m\times m}$ and $P_c \in \mathbb{R}^{n\times n}$, the row- and column-permuted matrix

$$h' = P_r h P_c \tag{8.10}$$

has exactly the same singular values as h. In particular,

$$d_p(h, h') = 0 \tag{8.11}$$

for every p, making d invariant under row- and column-permutations of h.

If P is a permutation matrix, then $P^\top P = I$, *i.e.* it is orthogonal. Furthermore, the singular values of any matrix h are given by the square root of the eigenvalues of $h^\top h$. If $h' = P_r\, h\, P_c$, then

$$(h')^\top (h') = (P_r\, h\, P_c)^\top \; (P_r\, h\, P_c) \tag{8.12}$$

$$= P_c^\top\, h^\top\, P_r^\top\, P_r\, h\, P_c \tag{8.13}$$

$$= P_c^\top\, h^\top\, h\, P_c \,. \tag{8.14}$$

Since $P_c^\top\, h^\top h\, P_c$ is a similarity transform of $h^\top h$, which does not change the eigenvalues, $h^\top h$ and $(h')^\top h'$ have the same eigenvalues, and in turn h and h' share the same singular values. hence $\sigma(h') = \sigma(h)$, and therefore

$$d_p(h, h') = \|\sigma(h) - \sigma(h')\|_p = 0 \,, \tag{8.15}$$

proving that, for any row or column permutation of h, the distance $d(h, h')$ remains unchanged.

8.5.2 Proof of Equivariance of Multi-Head Attention to Structured Permutations 8.2.1

We provide a detailed, step-by-step proof showing that our two-stage alignment procedure—inter-head reordering followed by intra-head permutations—preserves the functionality of a multi-head self-attention layer. Let:

- $X \in \mathbb{R}^{S\times d_{\text{model}}}$ be the input sequence.

- $W_q, W_k, W_v \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ be the query, key, and value projection matrices.
- H be the number of attention heads, each of dimensionality $d_k = d_{\text{model}}/H$.

Define:

$$Q = XW_q, \quad K = XW_k, \quad V = XW_v, \tag{8.16}$$

and split them by head:

$$Q = [Q_1, Q_2, \ldots, Q_H], \quad Q_i \in \mathbb{R}^{S \times d_k}, \tag{8.17}$$

and similarly for K and V. Let P_{inter} be an inter-head permutation in $\mathcal{S}_H$, with induced permutation vector π, and let $P_{\text{intra}}^{(i)} \in \mathcal{S}_{d_k}$ be the intra-head permutation for head i. We form the block-permutation matrix:

$$P_{\text{attn}} = \sum_{i=1}^{H} E^{i,\pi(i)} \otimes P_{\text{intra}}^{(i)}, \tag{8.18}$$

where $E^{i,\pi(i)}$ is a binary $H \times H$ matrix with a single 1 at $(i, \pi(i))$, and $\otimes$ denotes the Kronecker product.

Step 1: Permuting the projection weights

Applying P_{attn} to the query projections gives:

$$\begin{aligned} Q' = XW_qP_{\text{attn}} &= QP_{\text{attn}} \\ &= \left[\sum_{j=1}^{H} Q_j P_{\text{attn}}[j,i]\right]_{i=1}^{H} \\ &= \left[Q_{\pi^{-1}(i)} P_{\text{intra}}^{\pi^{-1}(i)}\right]_{i=1}^{H} \end{aligned}$$

Hence,

$$Q'_i = Q_{\pi^{-1}(i)}\, P_{\text{intra}}^{\pi^{-1}(i)}, \tag{8.19}$$

where the new head Q'_i corresponds to the head designated by the inter-head permutation $\pi^{-1}(i)$, modified according to $P_{\text{intra}}^{\pi^{-1}(i)}$. The same applies to:

$$K'_i = K_{\pi^{-1}(i)} P_{\text{intra}}^{\pi^{-1}(i)}, \quad V'_i = V_{\pi^{-1}(i)} P_{\text{intra}}^{\pi^{-1}(i)}. \tag{8.20}$$

Step 2: Permuting the attention scores

Because each $P_{\text{intra}}^{(i)}$ is orthogonal ($PP^T = I$), the attention scores satisfy:

$$\begin{aligned} A_i' &= \text{softmax}\left(\frac{Q_i' {K_i'}^T}{\sqrt{d_k}}\right) \\ &= \text{softmax}\left(\frac{Q_{\pi^{-1}(i)} P_{\text{intra}}^{\pi^{-1}(i)} (P_{\text{intra}}^{\pi^{-1}(i)})^T K_{\pi^{-1}(i)}^T}{\sqrt{d_k}}\right) \\ &= \text{softmax}\left(\frac{Q_{\pi^{-1}(i)} K_{\pi^{-1}(i)}^T}{\sqrt{d_k}}\right) = A_{\pi^{-1}(i)}. \end{aligned}$$

Thanks to the orthogonality of the intra-head permutation blocks, the attention scores are only influenced by the inter-head permutation.

Step 3: Permuting the value outputs

For each head,

$$O_i' = A_i' V_i' = A_{\pi^{-1}(i)} V_{\pi^{-1}(i)} P_{\text{intra}}^{\pi^{-1}(i)} = O_{\pi^{-1}(i)} P_{\text{intra}}^{\pi^{-1}(i)}. \tag{8.21}$$

Step 4: Reconstructing the final output

Concatenating all heads yields:

$$O' = [O_1', O_2', \ldots, O_H'] = O P_{\text{attn}}. \tag{8.22}$$

Conclusion. Applying P_{attn} to the projection matrices is thus equivalent to permuting the output of multi-head attention. The self-attention layer remains functionally equivalent, and the original output can be recovered via $O = O' P_{\text{attn}}^T$.

8.5.3 Full Procedure to Manage Residual Connections

We begin with the standard formulation of a transformer block, ignoring LayerNorm for simplicity:

$$\begin{aligned} \mathbf{z}_{attn} &= W_0 \,\text{MHA}(\mathbf{x}), \\ \mathbf{z}_i &= \mathbf{z}_{attn} + \mathbf{x}, \\ \mathbf{z}_f &= W_2 \text{ReLU}(W_1 \mathbf{z}_i), \\ \mathbf{z}_{out} &= \mathbf{z}_f + \mathbf{z}_i. \end{aligned} \tag{8.23}$$

Ignoring the ReLU activation function as well, we examine the impact of applying a permutation to one layer within a transformer block and then reversing it in the

subsequent layer. This transformation leads to:

$$\begin{aligned}
\mathbf{z}_{\text{attn}} &= P_{\text{W}_0} W_0 P_{\text{attn}}^{\top} \Big(P_{\text{attn}} \, \text{MHA} \, P_{\text{in}}^{\top} (P_{\text{in}} \mathbf{x}) \Big), \\
\mathbf{z}_i &= P_{\text{W}_0} \mathbf{z}_{\text{attn}} + P_{\text{in}} \mathbf{x}, \\
\mathbf{z}_f &= P_{\text{W}_2} W_2 P_{\text{W}_1}^{\top} \Big(P_{\text{W}_1} W_1 P_{\text{W}_0}^{\top} (\mathbf{z}_i) \Big), \\
\mathbf{z}_{\text{out}} &= P_{\text{W}_2} \mathbf{z}_f + P_{\text{W}_0} \mathbf{z}_i.
\end{aligned} \tag{8.24}$$

To ensure consistency in the permutation applied to both addends within each residual block, we replace the identity mapping with a permutation composition $\mathcal{I}$, where $\mathcal{I}_i = P_{\text{W}_0} P_{\text{in}}^{\top}$ and $\mathcal{I}_{\text{out}} = P_{\text{W}_2} P_{\text{W}_0}^{\top}$. This results in:

$$\begin{aligned}
\mathbf{z}_i &= P_{\text{W}_0} \mathbf{z}_{\text{attn}} + \mathcal{I}_i P_{\text{in}} \mathbf{x} = P_{\text{W}_0} \mathbf{z}_{\text{attn}} + P_{\text{W}_0} \mathbf{x}, \\
\mathbf{z}_{\text{out}} &= P_{\text{W}_2} \mathbf{z}_f + \mathcal{I}_{\text{out}} P_{\text{W}_0} \mathbf{z}_i = P_{\text{W}_2} \mathbf{z}_f + P_{\text{W}_2} \mathbf{z}_i.
\end{aligned} \tag{8.25}$$

After incorporating these compositions, the permutations remain consistent across each residual path, simplifying the block equations to:

$$\begin{aligned}
\mathbf{z}_{\text{attn}} &= P_{\text{W}_0} W_0 \Big(\text{MHA}(\mathbf{x}) \Big), \\
\mathbf{z}_i &= P_{\text{W}_0} \mathbf{z}_{\text{attn}} + P_{\text{W}_0} \mathbf{x}, \\
\mathbf{z}_f &= P_{\text{W}_2} W_2 \Big(W_1 (\mathbf{z}_i) \Big), \\
\mathbf{z}_{\text{out}} &= P_{\text{W}_2} \mathbf{z}_f + P_{\text{W}_2} \mathbf{z}_i.
\end{aligned} \tag{8.26}$$

With P_{W_2} serving as the input permutation for the subsequent layer.

8.5.4 Proof of Proposition 8.2.2

Proposition 8.5.2 *To assess how computational complexity scales with model size, we define:*

- *L: number of layers, evenly divided into MLP ($\frac{L}{2}$) and self-attention ($\frac{L}{2}$).*
- *H: number of attention heads.*
- *Each MLP layer contains two linear projections with dimension (d_m, d_h) and (d_h, d_m), assuming $d_m = d_h$.*
- *Self-attention layers have Q, K, and V matrices, each of size (d_m, d_m).*

We now estimate the complexity for a single iteration of the weight-matching algorithm.

MLP Layers. *The main computational cost comes from computing a pairwise similarity matrix between rows of projection matrices ($O(d_m^3)$), and solving a (d_m, d_m) assignment via Hungarian algorithm ($O(d_m^3)$). Hence, per-layer cost is:*

$$O(d_m^3) \tag{8.27}$$

Self-Attention Layers. *Split into two steps: inter-head and intra-head permutation.*

Inter-head permutation:

- $6H$ *SVDs over matrices of size* $(\frac{d_m}{H}, d_m)$*:*

$$O\left(\frac{6d_m^3}{H}\right) \tag{8.28}$$

- *Distance matrix over heads:*

$$O\left(\frac{3H^2 d_m}{2}\right) \tag{8.29}$$

- *Hungarian algorithm over* (H, H) *matrix:*

$$O(H^3) \tag{8.30}$$

Intra-head permutation:

- *Per-head similarity:*

$$O\left(\left(\frac{d_m}{H}\right)^2 d_m\right) = O\left(\frac{d_m^3}{H^2}\right) \tag{8.31}$$

- *Hungarian algorithm per head:*

$$O\left(\left(\frac{d_m}{H}\right)^3\right) \tag{8.32}$$

Summed over H *heads:*

$$O\left(H\left(\frac{d_m^3}{H^2} + \left(\frac{d_m}{H}\right)^3\right)\right) = O\left(\frac{d_m^3}{H} + \frac{d_m^3}{H^2}\right) \tag{8.33}$$

Total Self-Attention Cost per Layer.

$$O\left(\frac{6d_m^3}{H} + \frac{3H^2 d_m}{2} + H^3 + \frac{d_m^3}{H} + \frac{d_m^3}{H^2}\right) \tag{8.34}$$

Final Complexity. *Summing across* $\frac{L}{2}$ *MLP and* $\frac{L}{2}$ *attention layers:*

$$O\left(\frac{L}{2} d_m^3 + \frac{L}{2}\left(\frac{6d_m^3}{H} + \frac{3H^2 d_m}{2} + H^3 + \frac{d_m^3}{H} + \frac{d_m^3}{H^2}\right)\right). \tag{8.35}$$

This expression can be algebraically simplified to a more compact equivalent form:

$$O\left(L\left(d_m^3 + \frac{d_m^3}{H} + \frac{d_m^3}{H^2} + H^3 + H^2 d_m\right)\right). \tag{8.36}$$

So, the complexity scales polynomially with d_m and H, and remains significantly lower than data-based fine-tuning.

9. The Role of Pre-training for Model Merging in 3D Medical Segmentation

9.1 Overview

Although deep networks have achieved significant success in lesion segmentation and disease diagnosis [1, 84], medical image segmentation still poses distinct challenges in obtaining high-quality annotated data. The scarcity of labeled data, due to the time-intensive nature of manual annotation and the variability in imaging protocols across institutions, makes it challenging to build robust models. As a result, fully annotated datasets are often unavailable at the outset of a project, and new diseases or segmentation classes may emerge later. In this respect, models deployed in real-world healthcare settings should ideally learn continuously while preserving previously acquired knowledge. A straightforward approach for integrating new knowledge involves retraining the model from scratch on an aggregated dataset that includes both past and newly available data. However, strict privacy and security regulations may prohibit the

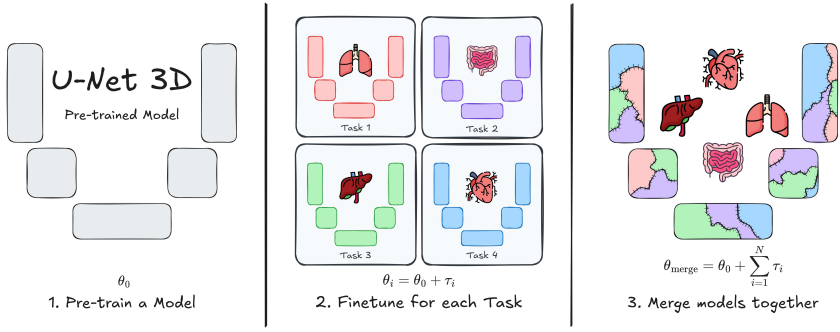


Figure 9.1: Overview of model merging for 3D medical segmentation models.

long-term storage of patient records, and resources for full retraining may be unavailable, making this approach impractical or undesirable in medical contexts.

To support these scenarios, an ideal model should allow for fast and flexible adaptation, enabling the integration of new data or classes. If the model can accommodate novel anatomical structures without requiring re-training, it would reduce storage and deployment costs and potentially reduce the need for labeled data. From a medical perspective, removing the need for complete re-training would minimize the long-term storage of sensitive training data, simplify compliance with ethical committee requirements, and support a decentralized and modular development paradigm. Commercially, the ability to combine model capabilities without re-training would enable dynamic, client-specific software customization, thereby accelerating deployment and offering greater flexibility. Notably, *model merging* permits updating and customizing AI models, facilitating knowledge transfer without full retraining [81, 120, 175, 199]. Our approach builds on these foundations by utilizing task vectors [81, 151], which represent modifications to a pre-trained model introduced during fine-tuning for a specific task. These vectors can be added to tune the model's functionalities (Fig. 9.1).

Unfortunately, model merging is not always practical, as it relies on the availability of effective pre-trained models. While standard computer vision tasks benefit from a wide selection of pre-trained base models, medical imaging—particularly tasks involving 3D segmentation—does not share the same advantage. In this respect, we aim to investigate the properties a base pre-trained model must possess to enable more effective model merging. Through both analytical and empirical assessments, we demonstrate why the base model should attain wide minima [200, 201] in the optimization landscape. While wide minima have been investigated in continual learning [125, 127] (*i.e.* tasks succeed one after the other), their implications in the context of model merging—where models are integrated simultaneously—remain unexplored until this study.

Specifically, we present the first analysis of model merging for 3D image segmentation. Considering two well-known medical datasets (ToothFairy2 [11, 12] and BTCV Abdomen [101]) and the standard 3D architecture, our study shows how models specialized for segmenting different anatomical structures can be successfully merged into a single model that can perform all the original tasks.

Contributions. We provide: *i)* an extensive analysis of model merging for 3D segmentation based on well-known medical datasets, revealing that combining task vectors is a flexible method for customizing models without re-training, *ii)* we offer both theoretical and empirical validation showing how a base model with a flat loss landscape enhances model merging, *iii)* alongside the source code, the model's weights are publicly released to facilitate research.

9.2 Framework

We deal with a neural network $f(\cdot;\boldsymbol{\theta})$ designed for 3D segmentation, such as . The model has weights $\boldsymbol{\theta} \in \mathbb{R}^m$ and takes 3D images as input $\boldsymbol{x} \in \mathbb{R}^{H\times W\times D}$. The output is a 3D map of class distributions $p_\theta(\boldsymbol{y}|\boldsymbol{x})$, one for each voxel $\boldsymbol{y}$ in $\mathcal{Y} \in \mathbb{R}^{H\times W\times D\times C}$. We study a **multi-task learning framework** comprising T segmentation tasks, denoted as $\mathcal{T}$. Each task $t \in \mathcal{T}$ is associated with a dataset $\mathcal{D}_t$ of n_t training samples, sampled from a task-specific distribution $p_t(\boldsymbol{x}, \boldsymbol{y})$. Despite variations in these distributions (*e.g.* different anatomical parts segmented in each task), all share a common loss function $\ell(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y})$ (*e.g.* the cross-entropy loss), defined as the negative log-likelihood $\ell(\boldsymbol{\theta}|\boldsymbol{x}, \mathcal{Y}) = -\sum_{\boldsymbol{y}\in\mathcal{Y}} \log p_\theta(\boldsymbol{y}|\boldsymbol{x})$.

Model Merging. To learn multiple segmentation tasks, we consider training a distinct set of weights for each task independently. We organize these models within a pool $\mathcal{P} = \{f(\cdot;\boldsymbol{\theta}_t) \mid \boldsymbol{\theta}_t \triangleq \boldsymbol{\theta}_0 + \boldsymbol{\tau}_t\}_{t\in\mathcal{T}}$ that can be expanded to accommodate for new tasks. Importantly, each model $f(\cdot;\boldsymbol{\theta}_t)$ is initialized from a shared set of **pre-trained weights** $\boldsymbol{\theta}_0$ and fine-tuned for its respective task. The displacement in weight space $\boldsymbol{\tau}_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}_0$ is called *task vector* [81] and, intuitively, it represents a direction in which the loss decreases for the t-th task.

As we discuss further, the models in the pool $\mathcal{P}$ can be selected and combined in arbitrary ways to construct a (personalized) multi-task model. The most straightforward approach to achieve this is by simply averaging the weights within the pool:

$$f_{\mathcal{P}} \triangleq f(\cdot;\boldsymbol{\theta}_{\mathcal{P}}) \quad \text{s.t.} \quad \boldsymbol{\theta}_{\mathcal{P}} = \boldsymbol{\theta}_0 + \textstyle\sum_{t=1}^{T} w_t\boldsymbol{\tau}_t, \quad \sum_{t=1}^{T} w_t = 1. \tag{9.1}$$

By adjusting the coefficients w_t, we can specialize the merged model for specific tasks while deprioritizing others. Conversely, for a model that maintains a balance across all tasks, a uniform weighting scheme, $w_t = 1/T$, can be used.

The **goal** is to design an approach that learns and combines multiple 3D segmentation models, ensuring the resulting merged model performs well across a set of combined tasks. To assess multi-tasking, we define the **empirical risk**, *i.e.* the average loss $\hat{\ell}(\boldsymbol{\theta}|\mathcal{D})$ over the union of all training tasks:[1]

$$\hat{\ell}(\boldsymbol{\theta}|\mathcal{D}) = \tfrac{1}{\sum_{t=1}^{T} n_t} \textstyle\sum_{\boldsymbol{x},\boldsymbol{y}\in\bigcup_{t=1}^{T}\mathcal{D}_t} \ell(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}) \tag{9.2}$$

Research Question. While 2D image classification tasks can benefit from a variety of pre-trained models (*e.g.* CLIP and DINO), 3D medical segmentation tasks face the absence of similar pre-trained models. In this respect, *how can we develop pre-trained models for 3D segmentation that facilitate model merging*?

[1] To simplify the notation, we will no longer explicitly denote the dependence of the loss on the data and write the individual loss and the empirical risk as $\ell(\boldsymbol{\theta})$ and $\hat{\ell}(\boldsymbol{\theta})$.

9.2.1 Model Merging from a Pre-training Perspective

Following [143], we analyze model merging through the lens of the Taylor approximation of the loss function. Specifically, we indicate as $\ell_{\text{cur}}(\boldsymbol{\theta})$ the second-order approximation of the empirical risk, centered around the pre-trained weights $\boldsymbol{\theta}_0$:

$$\hat{\ell}_{\text{cur}}(\boldsymbol{\theta}) = \hat{\ell}(\boldsymbol{\theta}_0) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{T}} \nabla \hat{\ell}(\boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathbf{H}_{\hat{\ell}}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0), \tag{9.3}$$

with $\nabla \hat{\ell}(\boldsymbol{\theta}_0) \triangleq \nabla_{\boldsymbol{\theta}} \hat{\ell}(\boldsymbol{\theta}_0)$ and $\mathbf{H}_{\hat{\ell}}(\boldsymbol{\theta}_0) \triangleq \nabla^2_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_0)$ indicating the gradient and the Hessian around $\boldsymbol{\theta}_0$. Based on [143], assuming that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is a local minimum for the empirical risk $\hat{\ell}(\boldsymbol{\theta})$ across all tasks, the Hessian is positive semi-definite. It follows that the second-order approximation $\hat{\ell}_{\text{cur}}(\boldsymbol{\theta})$ of the empirical risk is locally convex. Utilizing Jensen's inequality (valid for convex functions) we can establish the following relationship between the merged model and the individuals:

$$\hat{\ell}_{\text{cur}}(\boldsymbol{\theta}_{\mathcal{P}} = \boldsymbol{\theta}_0 + \textstyle\sum_{t=1}^{T} w_t \boldsymbol{\tau}_t) \leq \sum_{t=1}^{T} w_t \; \hat{\ell}_{\text{cur}}(\boldsymbol{\theta}_t = \boldsymbol{\theta}_0 + \boldsymbol{\tau}_t). \tag{9.4}$$

This inequality is informative because the term on the right provides a worst-case upper bound on the performance of the merged model. In particular, the empirical risk $\hat{\ell}_{\text{cur}}(\boldsymbol{\theta}_{\mathcal{P}})$ of the merged model is constrained by the convex combination of the empirical risks associated with each model. This implies that if each model $\boldsymbol{\theta}_t$ performs accurately across all tasks, there are certain assurances regarding the risk level of the merged model $\boldsymbol{\theta}_{\mathcal{P}}$.

However, the issue with Eq. 9.4 is that, under a scenario with specialized models trained on separate tasks, we cannot ensure that each model $\boldsymbol{\theta}_t$ performs well across all tasks. Indeed, as $\boldsymbol{\theta}_t$ is trained exclusively on its specific distribution $p_t(\boldsymbol{x}, \boldsymbol{y})$, its empirical risk is likely high for other data distributions $p_{t' \neq t}(\boldsymbol{x}, \boldsymbol{y})$ ($\rightarrow$ **low out-of-distribution performance**). For this reason, the following augmented optimization problem was proposed [143] for the t-th learner:

$$\underset{\boldsymbol{\theta}_t}{\text{minimize}} \quad \mathbb{E}_{\boldsymbol{x},\boldsymbol{y} \sim p_t(\boldsymbol{x},\boldsymbol{y})} \left[\ell_{\text{cur}}(\boldsymbol{\theta}_t | \boldsymbol{x}, \boldsymbol{y}) \right] + \mathcal{D}_{\text{KL}}(p_{\boldsymbol{\theta}_0}(\boldsymbol{y}|\boldsymbol{x}) || p_{\boldsymbol{\theta}_t}(\boldsymbol{y}|\boldsymbol{x})). \tag{9.5}$$

In essence, the out-of-distribution performance of each model is preserved through additional regularization provided by the term $\mathcal{D}_{\text{KL}}(\cdot)$, which acts explicitly on out-of-distribution examples $\boldsymbol{x}, \boldsymbol{y} \sim p_{t' \neq t}(\boldsymbol{x}, \boldsymbol{y})$. The $\mathcal{D}_{\text{KL}}(\cdot)$ term **aligns** the predictions $p_{\boldsymbol{\theta}_t}(\boldsymbol{y}|\boldsymbol{x})$ of the individual model $f(\cdot; \boldsymbol{\theta}_t)$ to those generated by the pre-trained model $\boldsymbol{\theta}_0$. By doing so, the individual model can achieve at least the performance level of the pre-trained model on external distributions $p_{t' \neq t}(\boldsymbol{x}, \boldsymbol{y})$, effectively reducing the upper bound on the right side of Eq. 9.4.

Table 9.1: Stable *vs.* plastic training regimes, metrics, and corresponding hyperparameters: Batch size (BS), Dropout (DO), and Learning rate (LR). λ_i correspond to the eigenvalues of $\mathbf{H}_{\hat{\ell}}(\boldsymbol{\theta}_0)$.

Regime	Dataset	BS	DO	LR	Dice↑	$\sum \boldsymbol{\lambda_i}\downarrow$	$\boldsymbol{\lambda_1}\downarrow$
Stable	Cui	4	0.5	10^{-3}	34.93	0.57	0.02
Plastic		8	0.0	10^{-4}	42.68	40.71	6.00
Stable	AMOS	4	0.5	10^{-3}	43.76	2.30	0.03
Plastic		8	0.0	10^{-4}	46.87	58.46	0.05

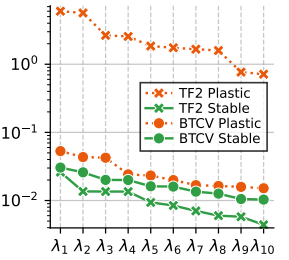


Figure 9.2: Top 10 eigenvalues ↓.

9.2.2 The Role of the Training Regime of the Pre-trained Model

While the authors of [143] drew inspiration from Eq. 9.5 to design a data-free regularization term, we take a different approach that avoids introducing explicit regularization. Instead, we focus on analyzing the roles of the training regime of the pre-trained model.

Thesis. We hypothesize that the tendency of the fine-tuned model $\boldsymbol{\theta}_t$ to retain pre-training knowledge is linked to the curvature of the pre-trained point $\boldsymbol{\theta}_0$ within the landscape of the empirical risk $\hat{\ell}(\cdot)$. To show that, we approximate the $\mathcal{D}_{\mathrm{KL}}(\cdot)$ as in [30]: if $\boldsymbol{\theta}_t - \boldsymbol{\theta}_0 \to 0$, the $\mathcal{D}_{\mathrm{KL}}(\cdot)$ term is close to the *distance* between $\boldsymbol{\theta}_t$ and the pre-training weights $\boldsymbol{\theta}_0$:

$$\mathcal{D}_{\mathrm{KL}}(p_{\boldsymbol{\theta}_0}(\boldsymbol{y}|\boldsymbol{x}) \;||\; p_{\boldsymbol{\theta}_t}(\boldsymbol{y}|\boldsymbol{x})) \approx \tfrac{1}{2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{H}_{\hat{\ell}}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_t - \boldsymbol{\theta}_0). \tag{9.6}$$

The weight distance is not isotropic but instead influenced by the Hessian of the empirical risk evaluated at $\boldsymbol{\theta}_0$. Thanks to Eq. 9.6 and the positive semi-definiteness of the Hessian around $\boldsymbol{\theta}_0$, we can establish a **bound** on $\mathcal{D}_{\mathrm{KL}}(\cdot)$:

$$\mathcal{D}_{\mathrm{KL}}(\dots) \approx \frac{1}{2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{H}_{\hat{\ell}}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_t - \boldsymbol{\theta}_0) \leq \frac{1}{2}\lambda_1 \left\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\right\|^2 = \frac{1}{2}\lambda_1 \left\|\boldsymbol{\tau}_t\right\|^2, \tag{9.7}$$

where λ_1 is the **maximum eigenvalue** of the Hessian $\mathbf{H}_{\hat{\ell}}(\boldsymbol{\theta}_0)$ around the pre-training weights. The result is that the degradation in out-of-distribution performance relative to the pre-trained model is controlled by: *i)* the norm of the task vector, and *ii)* the maximum eigenvalue λ_1 of the Hessian. Notably, the entire spectrum of eigenvalues has been crucial in analyzing the geometry of the loss landscape and its impact on generalization capabilities [44, 93]. Moreover, the maximum eigenvalue has been extensively used to characterize the width of a local minima [74, 93, 127]. In particular, a larger maximum eigenvalue suggests that the loss landscape is steeper along at least one dimension, which corresponds to a *sharper minimum*. Conversely, smaller eigenvalues suggest *wider minima* because the surface of the loss function changes less drastically in those

Table 9.2: Details of the datasets used in our experiments. Data is not resampled, but it is preprocessed with z-score normalization and patch-based training.

Dataset	Modality	Volumes	Structs	Shape
AMOS [87] (**pre-training**)	CT	240	15	$148 \times 533 \times 560$
BTCV Abdomen [101]		30	13	$125 \times 512 \times 512$
Cui [39] (**pre-training**)	CBCT	151	42	$322 \times 402 \times 402$
ToothFairy2 [12]		480	42	$169 \times 356 \times 375$

directions. Hence, to sum up, for a fixed task vector $\boldsymbol{\tau}_t$, the wider the curvature of the pre-trained model, the lower the loss in out-of-distribution performance during fine-tuning, and the better fine-tuned individual models will merge.

9.2.3 Biasing the Base Pre-Trained Model Towards Wide Minima

Building on this analytical finding, we propose modifying the training regime of the base pre-trained model to bias optimization toward wider minima. To do so, the approach is simple: inspired by [127], we act on some key hyperparameters—like batch size, dropout, and learning rate—that have been shown to affect generalization and the geometry of the minimum [56, 103, 194]. Following the terminology in [127], we define two distinct pre-training regimes, namely *stable* (wide minima) *vs. plastic* (sharp minima). The *stable* pre-training regime employs a small batch size, a higher learning rate, and increased dropout. In contrast, the *plastic* pre-training follows conventional self-supervised learning best practices, including using a large batch size, no dropout, and lower learning rates.

To analyze the effects of these hyperparameters, a preliminary result is reported in Tab. 9.1. We pre-train two base models (the one within the stable regime and the other in the plastic one) on two datasets for 3D medical image segmentation, namely AMOS [87] and Cui [39]. We then evaluate the average Dice on the corresponding test sets and compare the Hessian's eigenvalues as a proxy for the width of the pre-training optimum. Following [29], the Hessian's eigenspectrum is calculated with the trace of the empirical Fisher Information Matrix (FIM) [95], as a (diagonal) approximation of the intractable Hessian. As observed, the performance of the two base models (stable *vs.* plastic) is comparable across both datasets; however, the stable model achieves a remarkably lower trace (Fig. 9.2). This indicates that manipulating hyperparameters is a simple yet effective way to influence the geometry of the solution attained by the pre-trained model.

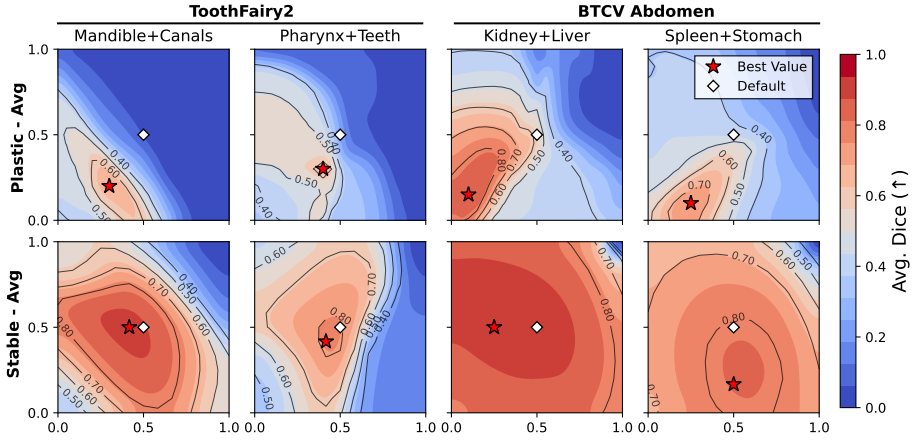


Figure 9.3: The average Dice score for two classes when merging task vectors τ_1 (x-axis) and τ_2 (y-axis) by varying w_1 and w_2. The star (★) marks the maximum Dice score, while the diamond (◇) denotes the default w_i values. The first row shows task vectors from *plastic* pre-training and the second row from *flat* pre-training, both merged using average. All the plots have the same x and y scales.

9.3 Experiments and Results

Datasets and Task Splits. Considering four public datasets, we categorize experiments into two settings based on the target anatomical regions: *i) abdominal datasets* (AMOS [87] and BTCV Abdomen [101]) and *ii) maxillofacial datasets* (Cui [39] and ToothFairy2 [12, 114]). The summary characteristics are provided in Tab. 9.2. In the abdominal scenario, we use AMOS for pre-training and four BTCV classes (Liver, Spleen, Kidney, and Stomach) to create four tasks. In the maxillofacial scenario, we use Cui for pre-training and ToothFairy2 for fine-tuning, with four tasks based on Mandible, Pharynx, Teeth, and Canals.

Training. We perform stable and plastic pre-training for both AMOS and Cui according to the setup in Tab. 9.1. To perform fine-tuning, we replace the final $1 \times 1 \times 1$ convolution with a new one; for the rest of the layers, we fine-tune the corresponding parameters $\boldsymbol{\theta}_0$ through a task vector $\boldsymbol{\tau}_t$ (initialized at zero). We optimize with *AdamW* [112] and a weight decay penalty of 0.1 to discourage large task vector norms. Training runs for 10 epochs.

9.3.1 Impact of Pre-Training Regime on Model Merging

In each plot of Fig. 9.3, we consider a pair of tasks (*e.g.* Mandible + Canals) and evaluate the Dice score of the merged model while varying merging coefficients w_1 and w_2, by comparing plastic (first row) *vs.* stable (second row) pre-training,

Table 9.3: Performance scores obtained from pairwise task vector merging.

Dataset	w	Merging Strategy	Spl. Kid.	Spl. Liv.	Spl. Sto.	Kid. Liv.	Kid. Sto.	Liv. Sto.	Avg.
BTCV Abdomen	Default ◇	Average [81]	91.41	92.18	80.61	90.85	77.18	80.91	85.52
		TIES [196]	82.80	90.76	76.83	88.69	58.56	76.69	79.05
	Best ★	Average [81]	92.64	92.22	82.09	91.01	78.97	81.80	86.45
		TIES [196]	92.42	91.88	81.55	91.07	77.72	81.04	85.95
	-	Joint	91.40	93.34	78.38	92.31	91.79	88.86	89.35

Dataset	w	Merging Strategy	Mand. IACs	Mand. Teeth	Mand. Phar.	IACs Teeth	IACs Phar.	Teeth Phar.	Avg.
ToothFairy2	Default ◇	Average [81]	89.54	82.55	87.70	56.08	57.08	81.27	75.70
		TIES [196]	88.70	79.89	88.96	58.51	63.44	73.72	75.54
	Best ★	Average [81]	89.67	82.78	91.89	68.16	75.19	81.27	81.49
		TIES [196]	89.24	82.33	91.97	67.94	68.91	81.07	80.24
	-	Joint	98.75	97.33	98.26	83.04	97.10	93.61	94.68

we can say that stable pre-training yields remarkably robust performance, exhibiting lower sensitivity to the merging coefficients — a feature that, in real-world applications, reduces the overhead associated with hyperparameter tuning. As further proof, for the stable regime, the uniform weighing scheme ◇ ($w_{1,2} = 0.5$) is always closer to the best configuration ★ (found by hyperparameter tuning on a held-out set).

After examining a scenario where pairs of tasks are merged, we extend our analysis to a setting with four task vectors. We report in Fig. 9.4 the results (Dice score) on each task separately and also the average (**Overall**). Beyond comparing stable *vs.* plastic pre-training, we also examine their impact on TIES Merging [196], a well-established alternative to uniform averaging. The results in Fig. 9.4 show that the performance of the merged model is influenced by the type of pre-training rather than the merging method. This is evidenced by the performance gains achieved with stable pre-training (*e.g.* uniform averaging), yielding improvements of +18.60 on ToothFairy2 and +16.28 on BTCV.

Further Comparative Analysis. To assess the effectiveness of model merging for 3D segmentation, we include a reference approach that re-trains from scratch, in which the pre-trained model is fine-tuned on both classes jointly. As shown in Tab. 9.3, in BTCV Abdomen, Kidney+Stomach shows the largest drop w.r.t. the joint training ($\sim$ 18.60 Dice score), while other pairs achieve similar performance, indicating effective merging. In contrast, the gap is significantly larger in ToothFairy2, likely due to greater variation in the shape, size, and intensity values of maxillofacial structures. We conjecture that increased variability leads to greater interference when merging relative task vectors.

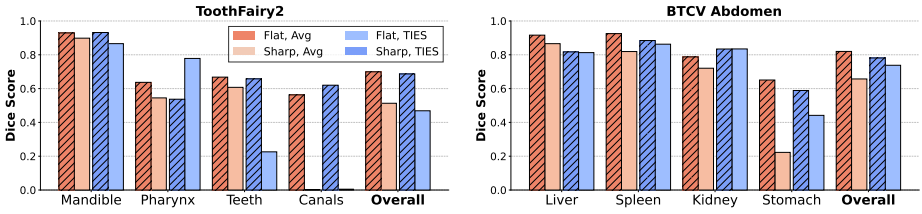


Figure 9.4: Task-wise performance after merging four distinct task vectors with weight averaging and TIES. The Overall bars aggregate results across tasks.

9.4 Discussion

In this work, we conduct a systematic study of model merging for 3D medical image segmentation and examine the influence of the base model's pre-training regime, which plays a central role. We show that the geometry of the solution reached during pre-training, in particular, the width of the minimum, strongly influences the quality of the task vectors and, consequently, the success of merging. Models trained under a *stable* regime, which encourages convergence toward flatter minima, consistently produce task vectors that combine more reliably across both abdominal and maxillofacial tasks.

Our findings reveal three key insights. First, task vectors obtained from stable pre-training yield merged models that are markedly less sensitive to merging coefficients, reducing the need for extensive hyperparameter tuning in practical deployments. Second, the benefits of stable pre-training persist regardless of the merging strategy: while methods such as TIES can improve performance, the dominant factor remains the curvature of the pre-trained optimum. Third, merging remains competitive with joint training in several scenarios, especially when anatomical variability is moderate, suggesting that model merging can serve as a viable alternative to full retraining, particularly in contexts where privacy constraints or computational limitations limit data access.

Taken together, these results point to a revised model life cycle for medical imaging, in which modular, data-efficient adaptation becomes feasible even in the absence of large-scale 3D pre-trained backbones.

Book IV

Conclusion

This thesis investigated three interconnected challenges about robustness and generalization in deep learning: the tendency of models to rely on shortcut solutions induced by explicit and implicit biases, the problem of catastrophic forgetting under spurious correlations, and the limited ability of independently trained models to transfer and compose knowledge.

In the first part, we addressed shortcut learning by introducing **ClusterFix**, a framework that leverages self-supervised clustering to identify and mitigate spurious correlations without requiring explicit protected-group annotations. We then extended this line of inquiry to continual learning through **Learning without Shortcuts (LwS)**, demonstrating that conventional rehearsal-based strategies are often insufficient to preserve robustness over time and may even exacerbate spurious correlations. To overcome this limitation, we proposed loss-based buffer sampling mechanisms that preserve worst-group performance across sequential tasks. Finally, we examined implicit biases in multimodal vision–language models, showing that CLIP-like architectures encode human-analogous biases and highlighted limitations of prompt-based debiasing approaches.

In the second part, the focus shifted to knowledge transfer and model composition in parameter space. We introduced **Transfusion**, a permutation-based alignment strategy that enables data-free transfer of task vectors across Transformer models trained on heterogeneous datasets. Building on this, we showed that the local geometry of the loss landscape plays a central role in determining merge compatibility. In particular, empirical evidence from 3D medical image segmentation showed that flatter pre-trained backbones yield more stable results when merging multiple fine-tuned models.

Modern deep learning systems increasingly depend on retraining or fine-tuning steps to accommodate new data, tasks, or architectural updates. While effective, this paradigm entails computational costs and becomes progressively less sustainable. The topics explored in this thesis align with a broader research direction that enables more efficient learning mechanisms, minimizing unnecessary recomputation while preserving performance and robustness. Within this perspective, learning is not a one-off optimization process, but an ongoing process of refinement and integration over time.

In conclusion, generalization in artificial intelligence should be understood as a dynamic system-level property, and a central open challenge is the development of AI systems that can evolve efficiently alongside their operating environments.

Appendix

Statement of contributions

For each of the works presented in this thesis, the following is a list of the contributions made by the candidate. The contributions are listed in the order in which they appear in the thesis.

- The research presented in Chap. 4 as **first author**. The candidate was responsible for the conceptualization of the debiasing approach, the design and execution of all experiments, the analysis of results, the preparation of all figures and visualizations, and contributed to the writing of the manuscript in collaboration with the co-authors.

- The research presented in Chap. 5 as **first author**. The candidate was responsible for the conceptualization of the method to handle spurious correlations in continual learning settings, the design and execution of all experiments, the analysis of results, the preparation of all figures and visualizations, and contributed to the writing of the manuscript in collaboration with the co-authors.

- The research presented in Chap. 6 as **first author**. The candidate was responsible for the conceptualization of the comprehensive bias analysis framework for vision-language models, the implementation of the evaluation pipeline, the execution of all experiments, the analysis of results across multiple bias dimensions, the preparation of all figures and visualizations, and contributed to the writing of the manuscript in collaboration with the co-authors.

- The research presented in Chap. 8 as **co-first author** (with Filippo Rinaldi). The candidate contributed to the conceptualization of the re-basin approach for task vectors, the implementation of core components of the method, the design and execution of experiments, the analysis of results, and contributed to the writing of the manuscript. The work was conducted in equal collaboration with the other co-first author.

- The research presented in Chap. 9 as **co-first author** (with Luca Lumetti). The candidate contributed to the conceptualization of the model merging framework for medical image segmentation, the implementation of core components of the method, the design and execution of experiments, the analysis of results, and contributed to the writing of the manuscript. The work was conducted in equal collaboration with the other co-first author.

List of publications

The following lists all works published in conference proceedings during the candidate's Ph.D. as a first author or first co-author:

Conference Proceedings

- "ClusterFix: A Cluster-Based Debiasing Approach without Protected-Group Supervision" **Giacomo Capitani**, Federico Bolelli, Angelo Porrello, Simone Calderara, Elisa Ficarra. IEEE/CVF Winter Conference on Applications of Computer Vision, 2024.

- "Beyond the Surface: Comprehensive Analysis of Implicit Bias in Vision-Language Models" **Giacomo Capitani**, Alice Lucarini, Lorenzo Bonicelli, Federico Bolelli, Simone Calderara, Loris Vezzali, Elisa Ficarra. ECCV 2024 Workshop (Fairness and Ethics towards transparent AI: facing the chalLEnge through model Debiasing), 2024.

- "Towards Unbiased Continual Learning: Avoiding Forgetting in the Presence of Spurious Correlations" **Giacomo Capitani**, Lorenzo Bonicelli, Angelo Porrello, Federico Bolelli, Simone Calderara, Elisa Ficarra. IEEE/CVF Winter Conference on Applications of Computer Vision, 2025.

- "Update Your Transformer to the Latest Release: Re-Basin of Task Vectors" Filippo Rinaldi, **Giacomo Capitani**, Lorenzo Bonicelli, Donato Crisostomi, Federico Bolelli, Elisa Ficarra, Emanuele Rodolà, Simone Calderara, Angelo Porrello. International Conference on Machine Learning, 2025

- "U-Net Transplant: The Role of Pre-training for Model Merging in 3D Medical Segmentation" Luca Lumetti, **Giacomo Capitani**, Elisa Ficarra, Simone Calderara, Costantino Grana, Angelo Porrello, Federico Bolelli. International Conference on Medical Image Computing and Computer Assisted Intervention, 2025

List of activities carried during the Ph.D.

Teaching activities

- Laboratory assistant, Fondamenti di Informatica (C Language), Department of Engineering Enzo Ferrari, University of Modena and Reggio Emilia, 2022-2023
- Laboratory assistant, Fondamenti di Informatica (C Language), Department of Engineering Enzo Ferrari, University of Modena and Reggio Emilia, 2023-2024

Advisor for master students

- Ethics in Artificial Intelligence Systems: CLIP Implicit Bias Analysis. Chiara Cappellino, 2023-2024.
- Mitigating Implicit Bias in Vision-Language Models: A CLIP Case Study. Alfredo Onori, 2023-2024.
- Structured Evaluation Strategies for Identifying Bias in Foundation Models. Katia Agrello, 2023-2024.

Participation to national and international projects

- **DECIDER**: DECIDER (Clinical Decision via Integrating Multiple Data Levels to Overcome Chemotherapy Resistance in High-Grade Serous Ovarian Cancer). Funded by the European Union from 1.2.2021 to 31.7.2026 under Horizon 2020 research and innovation program under grant agreement N. 965193.

Dissemination activities

- Poster presentation of the work "Update Your Transformer to the Latest Release: Re-Basin of Task Vectors" at the 2025 ICML international conference, Vancouver, Canada.
- Poster presentation of the work "Towards Unbiased Continual Learning: Avoiding Forgetting in the Presence of Spurious Correlations" at the 2025 WACV international conference, Tucson, Arizona.
- Poster presentation of the work "Beyond the Surface: A Comprehensive Analysis of Implicit Bias in Vision-Language Models" at the 2024 ECCV workshop, Milan, Italy.
- Poster presentation of the work "ClusterFix: A Cluster-Based Debiasing Approach without Protected-Group Supervision" at the 2024 WACV international conference, Waikoloa, Hawaii.

- Participation at the 2023 ELLIS Summer School on Large-Scale AI, Modena, Italy.
- Participation at the 2023 International Computer Vision Summer School, Scicli, Italy.

Review activities

- European Conference on Computer Vision (ECCV), 2024.
- IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, 2025.
- International Conference on Machine Learning (ICML), 2025.
- Neural Information Processing Systems (NeurIPS), 2025.

Internship

- Research Internship at the BRIC Centre (Copenhagen, Fran Supek Lab), August–October 2025: reconstruction and analysis of fusion protein sequences from genomic coordinates, functional domain characterization, and driver/passenger classification.

Bibliography

[1] Ravi Aggarwal, Viknesh Sounderajah, Guy Martin, Daniel SW Ting, Alan Karthikesalingam, Dominic King, Hutan Ashrafian, and Ara Darzi. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ Digital Medicine*, 4(1), 2021. 77

[2] Samuel Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git Re-Basin: Merging Models modulo Permutation Symmetries. In *International Conference on Learning Representations Workshop*, 2023. 52, 53, 54, 58, 59, 62, 65, 68

[3] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *European Conference on Computer Vision*, 2018. 11

[4] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. *Advances in neural information processing systems*, 32, 2019. 11, 26

[5] Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. Measuring implicit bias in explicitly unbiased large language models. *arXiv preprint arXiv:2402.04105*, 2024. 12, 13, 40

[6] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 2018. 33, 34

[7] Jihwan Bang, Hyunseo Koh, Seulki Park, Hwanjun Song, Jung-Woo Ha, and Jonghyun Choi. Online continual learning on a contaminated data stream with blurry task boundaries. In *IEEE conference on Computer Vision and Pattern Recognition*, 2022. 11

[8] John A Bargh, Mark Chen, and Lara Burrows. Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of personality and social psychology*, 71(2):230, 1996. 12, 40

[9] C Daniel Batson, Cynthia L Turk, Laura L Shaw, and Tricia R Klein. Information function of empathic emotion: Learning that we value the other's welfare. *Journal of personality and social psychology*, 68(2):300, 1995. 13

[10] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018. 9

[11] Federico Bolelli, Luca Lumetti, Shankeeth Vinayahalingam, Mattia Di Bartolomeo, Arrigo Pellacani, Kevin Marchesini, Niels Van Nistelrooij, Pieter Van Lierop, Tong Xi, Yusheng Liu, et al. Segmenting the Inferior Alveolar Canal in CBCTs Volumes: the ToothFairy Challenge. *IEEE Transactions on Medical Imaging*, 2024. ISSN 1558-254X. 78

[12] Federico Bolelli, Kevin Marchesini, Niels van Nistelrooij, Luca Lumetti, Vittorio Pipoli, Elisa Ficarra, Shankeeth Vinayahalingam, and Costantino Grana. Segmenting Maxillofacial Structures in CBCT Volumes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 78, 82, 83

[13] Lorenzo Bonicelli, Matteo Boschini, Angelo Porrello, Concetto Spampinato, and Simone Calderara. On the effectiveness of lipschitz-driven rehearsal in continual learning. *Advances in Neural Information Processing Systems*, 35:31886–31901, 2022. 26, 29

[14] Matteo Boschini, Lorenzo Bonicelli, Pietro Buzzega, Angelo Porrello, and Simone Calderara. Class-incremental continual learning into the extended der-verse. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 11

[15] Matteo Boschini, Pietro Buzzega, Lorenzo Bonicelli, Angelo Porrello, and Simone Calderara. Continual semi-supervised learning through contrastive interpolation consistency. *Pattern Recognition Letters*, 2022. 11

[16] Wieland Brendel and Matthias Bethge. Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. In *International Conference on Learning Representations*, 2019. 9

[17] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018. 4, 9

[18] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33: 15920–15930, 2020. 11, 26, 34, 35

[19] Pietro Buzzega, Matteo Boschini, Angelo Porrello, and Simone Calderara. Rethinking experience replay: a bag of tricks for continual learning. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2180–2187. IEEE, 2021. 26

[20] Lucas Caccia, Rahaf Aljundi, Nader Asadi, Tinne Tuytelaars, Joelle Pineau, and Eugene Belilovsky. New Insights on Reducing Abrupt Representation Change in Online Continual Learning. In *International Conference on Learning Representations Workshop*, 2022. 11, 34

[21] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. 41

[22] Giacomo Capitani, Federico Bolelli, Angelo Porrello, Simone Calderara, and Elisa Ficarra. Clusterfix: A cluster-based debiasing approach without protected-group supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4870–4879, 2024. 6, 10, 12, 27, 29, 32, 34, 35

[23] Giacomo Capitani, Alice Lucarini, Lorenzo Bonicelli, Federico Bolelli, Simone Calderara, Loris Vezzali, and Elisa Ficarra. Beyond the surface: A comprehensive analysis of implicit bias in vision-language models. In *European Conference on Computer Vision Workshops*, pages 35–52. Springer, 2024. 6, 13

[24] Giacomo Capitani, Lorenzo Bonicelli, Angelo Porrello, Federico Bolelli, Simone Calderara, and Elisa Ficarra. Towards unbiased continual learning: Avoiding forgetting in the presence of spurious correlations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2527–2537. IEEE Computer Society, 2025. 6, 11

[25] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 319–328, 2019. 14

[26] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018. 14

[27] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021. 46

[28] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018. 24

[29] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-SGD: Biasing Gradient Descent into Wide Valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12), 2019. 82

[30] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian Walk for Incremental Learning: Understanding Forgetting and Intransigence. In *European Conference on Computer Vision*, 2018. 81

[31] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. In *International Conference on Machine Learning Workshop*, 2019. 11

[32] Moon S Chen Jr, Primo N Lara, Julie HT Dang, Debora A Paterniti, and Karen Kelly. Twenty years post-nih revitalization act: Enhancing minority participation in clinical trials (empact): Laying the groundwork for improving minority clinical trial accrual: Renewing the case for enhancing minority participation in cancer clinical trials. *Cancer*, 120:1091–1096, 2014. 14

[33] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 41, 46, 65

[34] Leshem Choshen, Elad Venezian, Noam Slonim, and Yoav Katz. Fusing finetuned models for better pretraining. *arXiv preprint arXiv:2204.03044*, 2022. 53

[35] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023. 13, 41, 44, 45, 48

[36] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing Textures in the Wild. In *IEEE conference on Computer Vision and Pattern Recognition*, 2014. 65

[37] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013. 42

[38] Donato Crisostomi, Marco Fumero, Daniele Baieri, Florian Bernard, and Emanuele Rodola. C^2M^3: Cycle-Consistent Multi-Model Merging. *Advances in Neural Information Processing Systems*, 2024. 52

[39] Zhiming Cui, Yu Fang, Lanzhuju Mei, Bojun Zhang, Bo Yu, Jiameng Liu, Caiwen Jiang, Yuhang Sun, Lei Ma, Jiawei Huang, et al. A fully automatic AI system for tooth and alveolar bone segmentation from cone-beam CT images. *Nature Communications*, 13(1), 2022. 82, 83

[40] Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of data and analytics*, pages 296–299. Auerbach Publications, 2018. 4, 9

[41] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 10, 11

[42] Sepehr Dehdashtian, Lan Wang, and Vishnu Naresh Boddeti. Fairerclip: Debiasing clip's zero-shot predictions using functions in rkhss. *arXiv preprint arXiv:2403.15593*, 2024. 13

[43] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4, 9

[44] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp Minima Can Generalize For Deep Nets. In *International Conference on Machine Learning*, 2017. 81

[45] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4, 9, 46, 67

[46] Rachel Lea Draelos and Lawrence Carin. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks. *arXiv e-prints*, pages arXiv–2011, 2020. 24

[47] Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially No Barriers in Neural Network Energy Landscape. In *International Conference on Machine Learning*, 2018. 52

[48] Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The Role of Permutation Invariance in Linear Mode Connectivity of Neural Networks. In *International Conference on Learning Representations Workshop*, 2022. 52, 53

[49] Sebastian Farquhar and Yarin Gal. Towards Robust Evaluations of Continual Learning. In *International Conference on Machine Learning Workshop*, 2018. 11

[50] Enrico Fini, Victor G Turrisi Da Costa, Xavier Alameda-Pineda, Elisa Ricci, Karteek Alahari, and Julien Mairal. Self-supervised models are continual learners. In *IEEE conference on Computer Vision and Pattern Recognition*, pages 9621–9630, 2022. 29

[51] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 10

[52] Susan T Fiske, Amy JC Cuddy, and Peter Glick. Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, 11(2):77–83, 2007. 4, 12, 40

[53] Chloë FitzGerald and Samia Hurst. Implicit bias in healthcare professionals: a systematic review. *BMC medical ethics*, 18(1):1–18, 2017. 12

[54] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. 56

[55] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear Mode Connectivity and the Lottery Ticket Hypothesis. In *International Conference on Machine Learning*, 2020. 52, 53

[56] Jonathan Frankle, David J. Schwab, and Ari S. Morcos. The Early Phase of Neural Network Training. In *International Conference on Learning Representations Workshop*, 2020. 82

[57] C. Daniel Freeman and Joan Bruna. Topology and Geometry of Half-Rectified Network Optimization. In *International Conference on Learning Representations Workshop*, 2017. 52

[58] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4), 1999. 10, 53

[59] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024. 46, 65

[60] Qiankun Gao, Chen Zhao, Bernard Ghanem, and Jian Zhang. R-dfcil: Relation-guided representation learning for data-free class incremental learning. In *European Conference on Computer Vision*, 2022. 11

[61] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs. *Advances in Neural Information Processing Systems*, 2018. 52

[62] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bygh9j09KX`. 9, 12, 27

[63] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 4, 9, 28

[64] Robert Geirhos, Kristof Meding, and Felix A Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. *Advances in Neural Information Processing Systems*, 33: 13890–13902, 2020. 9, 12

[65] Anthony G Greenwald and Mahzarin R Banaji. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, 102(1): 4, 1995. 12, 40, 41, 43

[66] Robert J Grissom and John J Kim. *Effect sizes for research: A broad practical approach.* Lawrence Erlbaum Associates Publishers, 2005. 42

[67] Kailash A Hambarde and Hugo Proenca. Information Retrieval: Recent Advances and Beyond. *IEEE Access*, 2023. 12

[68] David L Hamilton. Stereotyping and intergroup behavior: Some thoughts on the cognitive approach. In *Cognitive processes in stereotyping and intergroup behavior*, pages 333–353. Psychology Press, 2015. 13, 40

[69] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4, 9

[70] 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. 33, 46

[71] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7), 2019. 65

[72] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. *IEEE International Conference on Computer Vision*, 2021. 65

[73] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 29

[74] Sepp Hochreiter and Jürgen Schmidhuber. Flat Minima. *Neural Computation*, 9(1), 1997. 81

[75] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *IEEE conference on Computer Vision and Pattern Recognition*, 2019. 11

[76] David A Houston. Empathy and the self: Cognitive and emotional influences on the evaluation of negative affect in others. *Journal of personality and social psychology*, 59(5):859, 1990. 13

[77] Frederick M Howard, James Dolezal, Sara Kochanny, Jefree Schulte, Heather Chen, Lara Heij, Dezheng Huo, Rita Nanda, Olufunmilayo I Olopade, Jakob N Kather, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature communications*, 12(1):4423, 2021. 9

[78] Lingxiao Huang and Nisheeth Vishnoi. Stable and fair classification. In *International Conference on Machine Learning*, pages 2879–2890. PMLR, 2019. 14

[79] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 2021. 57

[80] Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights. *Advances in Neural Information Processing Systems*, 2022. 53

[81] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations Workshop*, 2023. 78, 79, 84

[82] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing Models with Task Arithmetic. In *International Conference on Learning Representations Workshop*, 2023. 4, 5, 52, 53, 57, 65

[83] Moritz Imfeld, Jacopo Graldi, Marco Giordano, Thomas Hofmann, Sotiris Anagnostidis, and Sidak Pal Singh. Transformer Fusion with Optimal Transport. In *International Conference on Learning Representations Workshop*, 2024. 52, 59, 66, 68

[84] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2), 2021. 77

[85] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging Weights Leads to Wider Optima and Better Generalization. *arXiv preprint arXiv:1803.05407*, 2018. 53

[86] Myeongho Jeon, Hyoje Lee, Yedarm Seong, and Myungjoo Kang. Learning without prejudices: Continual unbiased learning via benign and malignant forgetting. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=gfPUokHsW-`. 11, 12, 26, 32, 34, 35

[87] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, et al. AMOS: ALarge-Scale Abdominal Multi-Organ Benchmark for Versatile Medical Image Segmentation. 35, 2022. 82, 83

[88] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021. 24

[89] Roy Jonker and Ton Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. In *Papers of the 16th Annual Meeting of DGOR in Cooperation with NSOR/Vorträge der 16. Jahrestagung der DGOR zusammen mit der NSOR*, 1988. 55

[90] Keller Jordan, Hanie Sedghi, Olga Saukh, Rahim Entezari, and Behnam Neyshabur. REPAIR: REnormalizing Permuted Activations for Interpolation Repair. In *International Conference on Learning Representations Workshop*, 2023. 52

[91] Irena Jovanović and Zoran Stanić. Spectral distances of graphs. *Linear Algebra and its Applications*, 436(5), 2012. 59

[92] Zhiqi Kang, Enrico Fini, Moin Nabi, Elisa Ricci, and Karteek Alahari. A soft nearest-neighbor framework for continual semi-supervised learning. In *IEEE International Conference on Computer Vision*, 2023. 11

[93] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In *International Conference on Learning Representations Workshop*, 2017. 81

[94] Chris Dongjoo Kim, Jinseo Jeong, Sangwoo Moon, and Gunhee Kim. Continual learning on noisy data streams via self-purified replay. In *IEEE International Conference on Computer Vision*, 2021. 11

[95] 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. 11, 82

[96] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 world wide web conference*, pages 853–862, 2018. 14

[97] Natasa Krco, Thibault Laugel, Jean-Michel Loubes, and Marcin Detyniecki. When mitigating bias is unfair: A comprehensive study on the impact of bias mitigation algorithms. *arXiv preprint arXiv:2302.07185*, 2023. 4, 14

[98] Alex Krizhevsky, Geoffrey Hinton, et al. Learning Multiple Layers of Features from Tiny Images. *Technical Report, University of Toronto*, 2009. 67

[99] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning*, pages 5905–5914. PMLR, 2021. 56

[100] Daniël Lakens. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas. *Frontiers in psychology*, 4:863, 2013. 42

[101] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial

vault–workshop and challenge. In *MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, 2015. 78, 82, 83

[102] Donggyu Lee, Sangwon Jung, and Taesup Moon. Continual learning in the presence of spurious correlations: Analyses and a simple baseline. In *International Conference on Learning Representations Workshop*, 2024. 11, 32, 34, 35, 39

[103] Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020. 82

[104] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 55(9):1–46, 2023. 9, 10

[105] Peizhao Li, Han Zhao, and Hongfu Liu. Deep fair clustering for visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9070–9079, 2020. 14

[106] Zhiheng Li, Anthony Hoogs, and Chenliang Xu. Discover and mitigate unknown biases with debiasing alternate networks. In *European Conference on Computer Vision*, pages 270–288. Springer, 2022. 10

[107] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021. 27, 32

[108] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 20

[109] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 33, 34

[110] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018. 26

[111] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017. 26

[112] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations Workshop*, 2019. 83

[113] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30, 2024. 57

[114] Luca Lumetti, Vittorio Pipoli, Federico Bolelli, Elisa Ficarra, and Costantino Grana. Enhancing Patch-Based Learning for the Segmentation of the Mandibular Canal. *IEEE Access*, 2024. 83

[115] Luca Lumetti, Giacomo Capitani, Elisa Ficarra, Simone Calderara, Costantino Grana, Angelo Porrello, and Federico Bolelli. U-net transplant: The role of pre-training for model merging in 3d medical segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 618–628. Springer, 2025. 7

[116] Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, et al. Fairclip: Harnessing fairness in vision-language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12289–12301, 2024. 13

[117] Utkarsh Mall, Cheng Perng Phoo, Meilin Kelsey Liu, Carl Vondrick, Bharath Hariharan, and Kavita Bala. Remote Sensing Vision-Language Foundation Models without Annotations via Ground Remote Alignment. In *International Conference on Learning Representations Workshop*, 2024. 57

[118] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *IEEE conference on Computer Vision and Pattern Recognition*, 2018. 11

[119] Arjun K Manrai, Birgit H Funke, Heidi L Rehm, Morten S Olesen, Bradley A Maron, Peter Szolovits, David M Margulies, Joseph Loscalzo, and Isaac S Kohane. Genetic misdiagnoses and the potential for health disparities. *New England Journal of Medicine*, 375(7):655–665, 2016. 14

[120] Michael S Matena and Colin A Raffel. Merging Models with Fisher-Weighted Averaging. *Advances in Neural Information Processing Systems*, 35, 2022. 53, 78

[121] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989. 4, 5, 10, 26, 53

[122] Kenneth O McGraw and Seok P Wong. A common language effect size statistic. *Psychological bulletin*, 111(2):361, 1992. 41, 42

[123] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 24

[124] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021. 9

[125] Sanket Vaibhav Mehta, Darshan Patil, Sarath Chandar, and Emma Strubell. An Empirical Investigation of the Role of Pre-training in Lifelong Learning. *Journal of Machine Learning Research*, 24(214), 2023. 56, 78

[126] Monica Millunzi, Lorenzo Bonicelli, Angelo Porrello, Jacopo Credi, Petter N Kolm, and Simone Calderara. May the forgetting be with you: Alternate replay for learning with noisy labels. In *British Machine Vision Conference*, 2024. 11

[127] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Razvan Pascanu, and Hassan Ghasemzadeh. Understanding the Role of Training Regimes in Continual Learning. *Advances in Neural Information Processing Systems*, 2020. 56, 78, 81, 82

[128] Clara Na, Sanket Vaibhav Mehta, and Emma Strubell. Train flat, then compress: Sharpness-aware minimization learns more compressible models. *arXiv preprint arXiv:2205.12694*, 2022. 56

[129] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from Failure: Training Debiased Classifier from Biased Classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020. 10, 12, 22, 31, 32, 34

[130] Aviv Navon, Aviv Shamsian, Idan Achituve, Ethan Fetaya, Gal Chechik, and Haggai Maron. Equivariant Architectures for Learning in Deep Weight Spaces. In *International Conference on Machine Learning*, 2023. 52

[131] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading Digits in Natural Images with Unsupervised Feature Learning. In *Neural Information Processing Systems Workshops*. Granada, 2011. 65

[132] Safiya Umoja Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, 2018. 12

[133] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pages 151–159, 2020. 10, 14

[134] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. 12

[135] Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task Arithmetic in the Tangent Space: Improved Editing of Pre-Trained Models. *Advances in Neural Information Processing Systems*, 2024. 57

[136] Charles E Osgood. Semantic differential technique in the comparative study of cultures. *American anthropologist*, 66(3):171–200, 1964. 4, 13, 40

[137] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71, 2019. 10, 26

[138] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *IEEE conference on Computer Vision and Pattern Recognition*, 2019. 11

[139] Razvan Pascanu and Yoshua Bengio. Revisiting natural gradient for deep networks. *arXiv preprint arXiv:1301.3584*, 2013. 55

[140] Fidel A Guerrero Peña, Heitor Rapela Medeiros, Thomas Dubail, Masih Aminbeidokhti, Eric Granger, and Marco Pedersoli. Re-basin via implicit Sinkhorn differentiation. In *IEEE conference on Computer Vision and Pattern Recognition*, 2023. 52

[141] Francesco Ponzio, Giacomo Deodato, Enrico Macii, Santa Di Cataldo, and Elisa Ficarra. Exploiting "uncertain" deep networks for data cleaning in digital pathology. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1139–1143. IEEE, 2020. 12

[142] Alice B Popejoy and Stephanie M Fullerton. Genomics is failing on diversity. *Nature*, 538(7624):161–164, 2016. 14

[143] Angelo Porrello, Lorenzo Bonicelli, Pietro Buzzega, Monica Millunzi, Simone Calderara, and Rita Cucchiara. A second-order perspective on compositionality and incremental learning. *arXiv preprint arXiv:2405.16350*, 2024. 11, 53, 80, 81

[144] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4, 9, 12, 40, 46, 57, 65

[145] Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019. 10

[146] Harish Guruprasad Ramaswamy et al. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 983–991, 2020. 24

[147] Alexandre Rame, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, patrick gallinari, and Matthieu Cord. Diverse Weight Averaging for Out-of-Distribution Generalization. In *Advances in Neural Information Processing Systems*, 2022. 53

[148] Alexandre Ramé, Kartik Ahuja, Jianyu Zhang, Matthieu Cord, Léon Bottou, and David Lopez-Paz. Model Ratatouille: Recycling Diverse Models for Out-of-Distribution Generalization. In *International Conference on Machine Learning*, 2023. 53

[149] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. iCaRL: Incremental classifier and representation learning. In *IEEE conference on Computer Vision and Pattern Recognition*, 2017. 11

[150] Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. *arXiv preprint arXiv:1810.11910*, 2018. 26

[151] Filippo Rinaldi, Giacomo Capitani, Lorenzo Bonicelli, Donato Crisostomi, Federico Bolelli, Elisa Ficarra, Emanuele Rodolà, Simone Calderara, and Angelo Porrello. Update Your Transformer to the Latest Release: Re-Basin of Task Vectors. In *International Conference on Machine Learning*, 2025. 7, 78

[152] Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018. 9

[153] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. 11

[154] Shiori Sagawa, Pang Wei Koh, Tatsunori Hashimoto, and Percy Liang. Distributionally Robust Neural Networks. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=ryxGuJrFvS`. 10, 14, 16, 17, 18, 21, 22, 23, 27

[155] Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, et al. Extending the wilds benchmark for unsupervised adaptation. *arXiv preprint arXiv:2112.05090*, 2021. 12, 32, 33

[156] Hadi Salman, Saachi Jain, Andrew Ilyas, Logan Engstrom, Eric Wong, and Aleksander Madry. When does bias transfer in transfer learning? *arXiv preprint arXiv:2207.02842*, 2022. 11

[157] Swami Sankaranarayanan, Thomas Hartvigsen, Lauren Oakden-Rayner, Marzyeh Ghassemi, and Phillip Isola. Real world relevance of generative counterfactual explanations. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*. 12

[158] Daniel L Schacter. Implicit memory: History and current status. *Journal of experimental psychology: learning, memory, and cognition*, 13(3):501, 1987. 43

[159] Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on mnist. *arXiv preprint arXiv:1805.09190*, 2018. 10

[160] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 46

[161] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 46

[162] Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *International conference on machine learning*, pages 4528–4537. PMLR, 2018. 29

[163] 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. 24

[164] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Unsupervised Learning of Debiased Representations with Pseudo-Attributes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16742–16751, 2022. 10, 12, 15, 17, 20, 21, 22, 23, 29, 32, 34, 35

[165] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*, 2017. 4, 9

[166] Sidak Pal Singh and Martin Jaggi. Model Fusion via Optimal Transport. *Advances in Neural Information Processing Systems*, 2020. 52, 53, 59

[167] James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *IEEE conference on Computer Vision and Pattern Recognition*, 2023. 11

[168] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No Subclass Left Behind: Fine-Grained Robustness in Coarse-Grained Classification Problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020. 10, 15, 17, 19, 21, 22, 23, 29, 32

[169] Matthew Staib and Stefanie Jegelka. Distributionally robust deep learning as a generalization of adversarial training. In *NIPS workshop on Machine Learning and Computer Security*, volume 3, page 4, 2017. 14

[170] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The German Traffic Sign Recognition Benchmark: A multi-class classification competition. In *The 2011 international joint conference on neural networks*. IEEE, 2011. 65

[171] Walter G Stephan and Cookie White Stephan. Intergroup anxiety. *Journal of social issues*, 41(3):157–175, 1985. 13, 40

[172] G Stoica, D Bolya, J Bjorner, P Ramesh, T Hearn, and J Hoffman. ZipIt! Merging Models from Different Tasks without Training. In *International Conference on Learning Representations Workshop*, 2024. 52

[173] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 9, 10

[174] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015. 4, 9

[175] Derek Tam, Mohit Bansal, and Colin Raffel. Merging by Matching Models in Task Parameter Subspaces. *Transactions on Machine Learning Research*, 2024. 78

[176] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2): 64–73, 2016. 46

[177] Bruce Thompson. Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools*, 44(5):423–432, 2007. 42

[178] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017. 10

[179] Antonio Torralba and Alexei A Efros. Unbiased Look at Dataset Bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011. 9

[180] Gido M Van de Ven, Hava T Siegelmann, and Andreas S Tolias. Brain-inspired replay for continual learning with artificial neural networks. *Nature communications*, 2020. 11

[181] Gido M van de Ven, Tinne Tuytelaars, and Andreas S Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4:1185–1197, 2022. 11

[182] Joaquin Vanschoren. Meta-learning: A survey. *arXiv preprint arXiv:1810.03548*, 2018. 10

[183] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NIPS)*, 30, 2017. 4, 9

[184] Radu-Laurenţiu Vieriu, Sergey Tulyakov, Stanislau Semeniuta, Enver Sangineto, and Nicu Sebe. Facial Expression Recognition under a Wide Range of Head Poses. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–7. IEEE, 2015. 12

[185] Jeffrey S Vitter. Random sampling with a reservoir. *ACM Transactions on Mathematical Software (TOMS)*, 11(1):37–57, 1985. 26

[186] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 21

[187] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *International Conference on Learning Representations Workshop*, 2019. 66

[188] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 9

[189] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, 2022. 11

[190] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *IEEE conference on Computer Vision and Pattern Recognition*, 2022. 11

[191] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, 2022. 52, 53, 64

[192] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *IEEE conference on Computer Vision and Pattern Recognition*, 2022. 53, 64

[193] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *IEEE conference on Computer Vision and Pattern Recognition*, 2019. 11

[194] Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent escapes from sharp minima exponentially fast. *arXiv preprint arXiv:2002.03495*, 2020. 82

[195] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023. 46

[196] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. TIES-Merging: Resolving Interference When Merging Models. *Advances in Neural Information Processing Systems*, 36, 2024. 53, 84

[197] Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. Clusterfit: Improving generalization of visual representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6509–6518, 2020. 19

[198] Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. Clusterfit: Improving generalization of visual representations. In *IEEE conference on Computer Vision and Pattern Recognition*, June 2020. 29

[199] Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. Adamerging: Adaptive model merging for multi-task learning. In *International Conference on Learning Representations Workshop*, 2024. 78

[200] Zhewei Yao, Amir Gholami, Qi Lei, Kurt Keutzer, and Michael W. Mahoney. Hessian-based Analysis of Large Batch Training and Robustness to Adversaries. 2018. 78

[201] Fuxun Yu, Zhuwei Qin, Chenchen Liu, Liang Zhao, Yanzhi Wang, and Xiang Chen. Interpreting and Evaluating Neural Network Robustness. In *International Joint Conference on Artificial Intelligence*, 2019. 78

[202] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 18, 21

[203] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. PMLR, 2017. 11

[204] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018. 10

[205] Frederic Z Zhang, Paul Albert, Cristian Rodriguez-Opazo, Anton van den Hengel, and Ehsan Abbasnejad. Knowledge Composition using Task Vectors with Learned Anisotropic Scaling. *Advances in Neural Information Processing Systems*, 2024. 69

[206] Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*, 2022. 12, 32

[207] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding. *CoRR*, abs/1610.02055, 2016. URL `http://arxiv.org/abs/1610.02055`. 21

[208] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 9