Mitigating Negative Flips via Margin Preserving Training

Simone Ricci, Niccolò Biondi, Federico Pernici, Alberto Del Bimbo

DINFO (Department of Information Engineering), University of Florence, Italy
MICC (Media Integration and Communication Center)
<name>.<surname>@unifi.it

Abstract

Minimizing inconsistencies across successive versions of an AI system is as crucial as reducing the overall error. In image classification, such inconsistencies manifest as negative flips, where an updated model misclassifies test samples that were previously classified correctly. This issue becomes increasingly pronounced as the number of training classes grows over time, since adding new categories reduces the margin of each class and may introduce conflicting patterns that undermine their learning process, thereby degrading performance on the original subset. To mitigate negative flips, we propose a novel approach that preserves the margins of the original model while learning an improved one. Our method encourages a larger relative margin between the previously learned and newly introduced classes by introducing an explicit margin-calibration term on the logits. However, overly constraining the logit margin for the new classes can significantly degrade their accuracy compared to a new independently trained model. To address this, we integrate a double-source focal distillation loss with the previous model and a new independently trained model, learning an appropriate decision margin from both old and new data, even under a logit margin calibration. Extensive experiments on image classification benchmarks demonstrate that our approach consistently reduces the negative flip rate with high overall accuracy.

Code — https://github.com/miccunifi/negative_flip_MPT

Introduction

Recent advances in machine learning have required the frequent deployment of updated models in production systems (Raffel 2023; Yadav et al. 2025). These updates often introduce new models that leverage more expressive network architectures (Touvron et al. 2023), novel training techniques or paradigms (Biondi et al. 2024; Echterhoff et al. 2024; Shen et al. 2020; Zhou et al. 2023), and additional data (Gunasekar et al. 2023). However, replacing an existing model requires balancing the potential reduction in overall error rates against the risk of introducing new errors that can disrupt downstream processes (Milani Fard et al. 2016; Yan et al. 2021) or lead to unexpected system behaviors for human users (Bansal et al. 2019). This regression[1] phenomenon, observed in both visual and natural language tasks (Yan et al. 2021; Abdul Samadh et al. 2024), arises from the occurrence of negative flips—instances that are correctly classified by the old model but misclassified by the new one (see Figure 1).

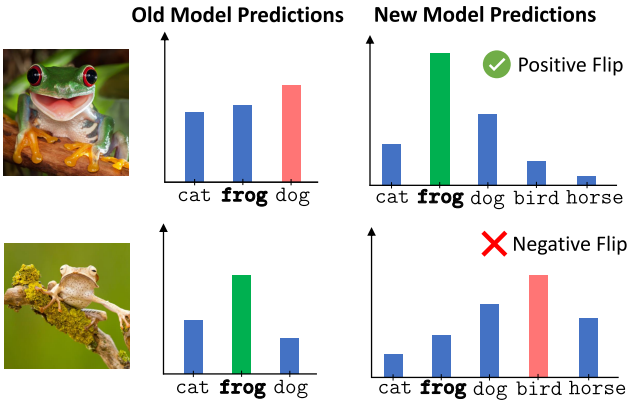

Figure 1: Illustration of possible two prediction changes following a model update. Each row shows the predictions of the old model (left) and the new model (right) for a given image, showing two possible cases: (1) a Positive Flip, where the old model was wrong and the new model is correct—the desirable outcome after a model update; (2) a Negative Flip, where the old model was correct but the new model is wrong, which can lead to unexpected system behaviors. In this paper, we focus on reducing Negative Flips in a classification task.

Negative flips in classification tasks are most likely to occur when the number of classes increases to accommodate new categories (Yan et al. 2021; Zhao et al. 2024; Zhang et al. 2021). From a margin-based perspective, increasing the number of classes at each model update reduces the margin between adjacent classes, due to class means pushing them-

[1]The term *regression*—from the software industry—denotes a decline in system performance following an update. Although an updated model may demonstrate improved average performance, any instances of regression can potentially disrupt downstream post-processing workflows.

selves to arrange in a way that maximizes the minimum one-vs-rest margin (Jiang et al. 2024a; Papyan, Han, and Donoho 2020; Yaras et al. 2022; Tirer, Huang, and Niles-Weed 2023; Pernici et al. 2019, 2021). In addition to the reduction imposed by the optimization process and geometric constraints (Jiang et al. 2024a), when new categories are introduced, the decision margins are subject to further challenges arising from negative transfer (Li, Nguyen, and Zhang 2023; Zhang et al. 2022). As the model learns to encode features that discriminate among an expanded set of classes, the incorporation of additional categories may lead to the emergence of interfering features, thereby adversely impacting the performance on previously acquired tasks or classes (Tirer, Huang, and Niles-Weed 2023; Liu et al. 2021; Torralba, Murphy, and Freeman 2007). In extreme cases, negative flips involve not only ambiguous samples near the decision boundaries but also high-confidence samples influenced, for example, by spurious correlations (Wang et al. 2021).

Despite these challenges, most existing approaches for reducing negative flips (Yan et al. 2021; Zhao et al. 2024; Zhang et al. 2021) overlook a key factor: *the progressive reduction of class margins as new categories are added*. Smaller margins make the model more susceptible to negative flips, since even confident, previously correct predictions can become incorrect due to slight shifts in decision boundaries. Therefore, preserving or restoring these margins is critical for maintaining consistent performance across model updates—especially for previously learned classes. This observation motivates us to focus on margin preservation as a central component of our approach. To this end, we preserve the logit decision margins[2] associated with previously learned classes, thereby addressing the challenge of negative flips when new classes are incrementally introduced. Specifically, we introduce a positive logit bias for the new classes as a margin-calibration term within the softmax cross-entropy loss, thereby modifying the optimization process such that the decision margins are biased towards the previously learned classes. Although this adjustment effectively reduces negative flips for previously learned classes, it may also result in underfitting of the newly introduced ones, as the model is encouraged to learn a smaller margin for them. To overcome this limitation and ensure appropriate margin learning for both previously learned and newly introduced classes, we propose a double-source focal distillation strategy. We distill knowledge corresponding to the correct predictions on the newly introduced categories from a model independently trained on the complete set of classes, while at the same time leveraging the previous model to preserve its correct behavior on the original categories.

Our main contributions are as follows:

- We present a novel approach, called Margin Preserving Training, to address the reduction of negative flips by focusing on the preservation of decision margins for previously learned classes.
- We propose a double-source focal distillation strategy to mitigate the negative impact of the margin-calibration term introduced in the training loss on newly added classes, thereby ensuring balanced and robust classification performance for both new and old categories.
- We conduct extensive experiments on image classification benchmarks, demonstrating that our method consistently reduces negative flips while maintaining or improving overall accuracy on the CIFAR100 and ImageNet1K datasets.

[2]The logit decision margin is the difference between the logit of the ground-truth class and the largest logit among all other classes. Following Pleiss et al. (2020) notation, the logit margin for an input sample of class y is defined as $\gamma(\mathbf{z}, y) = \mathbf{z}_y - \max_{j \neq y} \mathbf{z}_j$, where $\mathbf{z}$ denotes the logit vector from the final (pre-softmax) layer.

Related Works

Negative Flips Reduction. Negative flips, formalized by (Yan et al. 2021) expanding upon earlier notions of predictive churn (Milani Fard et al. 2016; Toneva et al. 2019), are related to the broader areas of continual learning (Chen and Liu 2018; Kirkpatrick et al. 2017), incremental learning (Li and Hoiem 2017; Prabhu, Torr, and Dokania 2020), and sequential learning (Goodfellow et al. 2013; McCloskey and Cohen 1989). While these areas focus on reducing forgetting and maintaining average performance across tasks (Delange et al. 2021; Kirkpatrick et al. 2017; Lopez-Paz and Ranzato 2017; Zenke, Poole, and Ganguli 2017), negative flips instead highlight instance-level prediction instability. To address this issue, most existing methods aim to align predictions across versions via ensemble techniques (Zhao et al. 2024; Xie et al. 2021) or knowledge distillation (Yan et al. 2021; Parchami-Araghi et al. 2024; Echterhoff et al. 2024). Distillation methods (Yan et al. 2021; Parchami-Araghi et al. 2024; Echterhoff et al. 2024; Träuble et al. 2021; Jiang et al. 2021) typically focus on output alignment but do not explicitly preserve decision margins, which can increase negative flips (Wu et al. 2022). For instance, Positive-congruent training (PCT) (Yan et al. 2021) proposes Focal Distillation that selectively distills only correct predictions from the old model, reducing the occurrence of negative flips. Ensemble methods reduce stochastic variations in training and improve class separation margins by averaging the outputs of multiple models (Zhao et al. 2024; Xie et al. 2021; Bahri and Jiang 2021; Zhang et al. 2021; Cai et al. 2022; Li et al. 2023), but incur a high cost at inference. To reduce deployment costs and mitigate negative flips, ELODI (Zhao et al. 2024) distills the ensemble knowledge to a single model. However, existing approaches either incur significant computational overhead or do not explicitly address the preservation of the decision margin when additional classes are introduced. For this reason, we propose a novel approach that overcomes both of these limitations.

Knowledge Distillation. Knowledge distillation (Hinton, Vinyals, and Dean 2015; Beyer et al. 2022) was originally developed to transfer knowledge from a larger teacher model to a smaller student. Since then, it has evolved to include advanced frameworks such as self-distillation (Zhang et al. 2019; Anil et al. 2018; Caron et al. 2021), teacher-ensemble strategies (Chebotar and Waters 2016; You et al. 2017; Malinin, Mlodozeniec, and Gales 2020; Asif, Tang, and Harrer 2020; Stanton et al. 2021), continual learning (Li and

Hoiem 2017; Asadi et al. 2023; Mistretta et al. 2024), and compatibility-aware approaches (Shen et al. 2020; Zhou et al. 2023; Biondi et al. 2023a,b; Ricci et al. 2024; Biondi et al. 2024; Ricci et al. 2025). Multi-source and multi-level distillation, leveraging several teachers or latent representations, have also been explored for greater robustness with applications in various contexts (Amirkhani et al. 2021; Yuan et al. 2021; Jiang et al. 2024b; Liu, Zhang, and Wang 2020). While knowledge distillation has been widely adopted to reduce negative flips (Yan et al. 2021; Zhao et al. 2024), existing approaches often rely on complex ensembles or single-base model distillation, which can be inefficient or limited, respectively. We propose a double-source focal distillation framework combined with decision margin preservation, which matches ensemble performance using only two models and also overcomes the limitations of single-model distillation by better preserving decision boundaries and reducing negative flips.

Margin Preservation. The concept of margins is foundational in a wide range of machine learning algorithms (Bartlett 1996; Bartlett et al. 1998; Weinberger and Saul 2009). Both empirical and theoretical work show that margins are closely connected to neural network generalization (Bartlett, Foster, and Telgarsky 2017; Elsayed et al. 2018; Jiang et al. 2019; Neyshabur et al. 2019), and play a key role in tasks such as classification (Pleiss et al. 2020), face recognition (Deng et al. 2019; Liu et al. 2017; Wang et al. 2018), long-tail distribution learning (Ren et al. 2020; Cao et al. 2019; Menon et al. 2021), and continual learning (Wu et al. 2019; Kirkpatrick et al. 2017). Recent work demonstrates that the final stage of classifier training can be seen as a margin optimization problem, directly linking the number of classes to the geometry of the learned representation (Jiang et al. 2024a). However, despite the recognized importance of margins for generalization and robustness, decision margin preservation has not been systematically studied in the context of negative flip reduction. Our work is the first to explicitly address this gap.

Problem Formulation

Given an image $\mathbf{x} \in \mathcal{X}$ and its ground-truth label $y \in \mathcal{Y}$, let q_ϕ denote the predicted probability distribution over classes given an input for a model parameterized by ϕ. The final predicted class for an input $\mathbf{x}_i$ is given by:

$$\hat{y}_i = \arg\max_{y \in \mathcal{Y}} q_\phi(y \mid \mathbf{x}_i).$$

Consider a base model ϕ_{old} trained on an initial set of classes $\mathcal{Y}^{\text{old}} = \{1, 2, \ldots, C^{\text{old}}\}$ and an improved version ϕ_{new}, which is trained on the expanded class set $\mathcal{Y}^{\text{all}} = \mathcal{Y}^{\text{old}} \cup \mathcal{Y}^{\text{new}}$, where $\mathcal{Y}^{\text{new}} = \{C^{\text{old}} + 1, \ldots, C^{\text{new}}\}$ represents the set of new classes. By analyzing the predictions of ϕ_{old} and ϕ_{new} on the original class set $\mathcal{Y}^{\text{old}}$, we can categorize different prediction outcomes based on $\hat{y}^{\text{old}}$ and $\hat{y}^{\text{new}}$. Predictions are considered consistent when both models either correctly classify a sample ($\hat{y}_i^{\text{old}} = \hat{y}_i^{\text{new}} = y_i$) or misclassify it ($\hat{y}_i^{\text{old}} \neq y_i \wedge \hat{y}_i^{\text{new}} \neq y_i$). However, in some cases, the predictions differ. A positive flip occurs when a sample previously misclassified by ϕ_{old} is correctly classified by ϕ_{new}. These flips are considered beneficial in model updates and should therefore be encouraged. Conversely, a negative flip occurs when a sample that was correctly classified by ϕ_{old} is misclassified by ϕ_{new}, making it a key factor in the phenomenon of performance regression. Figure 1 illustrates an example of both positive and negative flips.

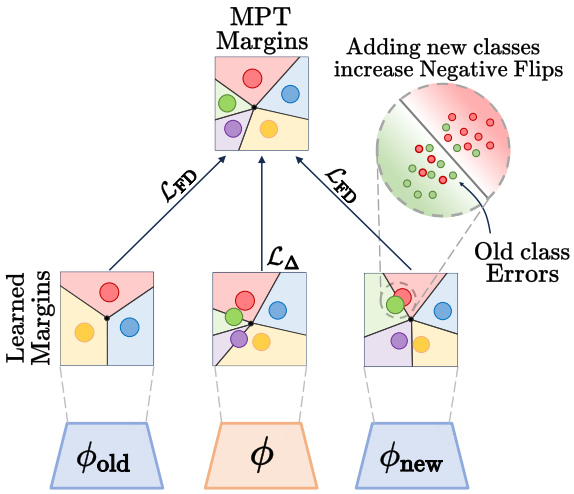

Figure 2: Schematic overview of the proposed Margin Preserving Training (MPT) approach. A model ϕ is trained with the margin-calibrated loss $\mathcal{L}_\Delta$ while receiving double-source focal distillation from two reference models, ϕ_{old} and ϕ_{new} (both optimized with cross-entropy). This combination mitigates margin underestimation for new classes, reduces negative flips between model updates, and preserves decision margins for old classes. The gray region marks areas near decision boundaries where negative flips typically occur. Margins are illustrated in a 2D embedding space for clarity, although MPT operates on logit-space margins over all classes.

Margin Preserving Training (MPT)

In this section, we introduce our Margin Preserving Training method, which aims to reduce negative flips between successive model updates. Our approach is characterized by (1) the incorporation of a bias term in the softmax cross-entropy loss (Equation 2) to preserve decision margins, and (2) the introduction of a double-source focal distillation term (Equation 3) to promote balanced learning between old and new classes. Figure 2 provides a schematic overview of our approach, illustrating the impact of each component on the learned class margins.

Margin-calibrated Softmax Cross-entropy Loss

When a model is updated by increasing the number of training classes, the margin between adjacent classes becomes narrower as they arrange themselves to maximize the minimum one-vs-rest margin (Jiang et al. 2024a). This contraction may lead to negative flips, particularly among instances of previously learned classes, as they become increasingly

Method	CIFAR100				ImageNet1k			
	ER$_\downarrow$(%) on $\mathcal{Y}^{\text{old}}$	ER$_\downarrow$(%) on $\mathcal{Y}^{\text{all}}$	NFR$_\downarrow$(%)	Rel-NFR$_\downarrow$(%)	ER$_\downarrow$(%) on $\mathcal{Y}^{\text{old}}$	ER$_\downarrow$(%) on $\mathcal{Y}^{\text{all}}$	NFR$_\downarrow$(%)	Rel-NFR$_\downarrow$(%)
Old model	33.27	-	-	-	25.55	-	-	-
No treatment	40.40	39.12	14.04	52.39	28.40	32.88	9.15	43.30
BCT (Shen et al. 2020)	41.28	39.58	14.88	54.34	27.90	33.17	8.99	43.28
PCT-Naive (Yan et al. 2021)	40.54	40.32	14.04	52.20	26.22	33.17	7.24	37.06
PCT-KL (Yan et al. 2021)	39.44	38.79	11.94	45.63	27.73	32.69	8.09	39.19
PCT-LM (Yan et al. 2021)	41.34	41.83	12.12	44.19	27.98	34.70	7.55	36.24
ELODI (Zhao et al. 2024)	37.98	38.45	10.56	41.91	27.28	**32.13**	7.61	37.46
ELODI$_{TopK}$ (Zhao et al. 2024)	39.06	**38.32**	12.46	48.09	28.51	32.97	9.07	42.74
MPT-KL (Ours)	**34.32**	38.55	**9.26**	40.67	**24.68**	32.76	6.09	33.15
MPT-LM (Ours)	35.28	38.61	**9.26**	**39.57**	25.46	34.12	**6.02**	**31.76**

Table 1: Comparison of different methods on CIFAR100 and ImageNet1k with a ResNet-18 architecture. We report the error rate on old classes (ER on $\mathcal{Y}^{\text{old}}$), the overall error rate (ER on $\mathcal{Y}^{\text{all}}$), the Negative Flip Rate (NFR), and the Relative Negative Flip Rate (Rel-NFR).

closer to each other. Furthermore, expanding the training set with new classes can introduce interfering features, thereby leading to negative transfer (Zhang et al. 2022). This phenomenon arises because the model tends to learn more generic rather than class-specific features (Li, Nguyen, and Zhang 2023), which increases the probability of negative flips. To address this, we propose a margin-calibrated softmax cross-entropy loss, designed to preserve the margins of previously learned classes. The core idea is to apply a positive logit bias to the newly introduced categories during training, tilting the decision boundary to favor of the old ones and thus maintaining their margin close to those learned by the base model, mitigating negative flips.

Given a logit vector $\mathbf{z}$ produced by a model ϕ for an input sample $\mathbf{x}$, where $\mathbf{z}_i$ corresponds to class i (with C the total number of classes), and y is the ground-truth class label, we introduce a class-dependent logit bias Δ_i. The logit bias Δ_i is defined as a positive bias $k > 0$ applied exclusively to the newly introduced classes $\mathcal{Y}^{\text{new}}$:

$$\Delta_i = \begin{cases} 0, & \text{if } i \in \mathcal{Y}^{\text{old}}, \\ k, & \text{if } i \in \mathcal{Y}^{\text{new}}. \end{cases} \tag{1}$$

The resulting margin-calibrated softmax cross-entropy loss is defined as:

$$\mathcal{L}_\Delta = -\log \frac{e^{\mathbf{z}_y + \Delta_y}}{\sum_{i=1}^{C} e^{\mathbf{z}_i + \Delta_i}}. \tag{2}$$

Introducing a positive bias into the logit functions serves as an explicit prior, increasing the predicted confidence for targeted classes during training. This approach reduces the model's dependency on highly discriminative features for those classes, as the introduced bias inherently supports correct classification. Conversely, classes without the positive logit bias must rely exclusively on feature-based discrimination, prompting the model to develop larger decision margins for these unbiased categories. At inference time, with the bias removed, the model's confidence in previously biased classes diminishes, revealing their comparatively weaker internal representations. Consequently, the absence of this prior forces the model to maintain wider decision margins for unbiased categories, as these posed greater challenges during training without the benefit of a positive bias (Ren et al. 2020). When the value of $k = 0$, then $\Delta_i = 0$ for all classes, and Equation 2 reduces to the standard cross-entropy loss, resulting in a balanced learning of margins across all categories.

Margin-calibrated losses have previously been formalized and employed to address class imbalance learning, where the most frequent classes receive larger margins due to disparities in class frequencies compared to rare ones (Ren et al. 2020; Menon et al. 2021; Cao et al. 2019). Motivated by this, we reinterpret the class-dependent logit bias as a tool for more general margin regulation. In our approach, Δ_i is not determined by class frequencies, as is common in class imbalance methods, but is instead governed by a tunable hyperparameter k that directly controls the trade-off between preserving old class margins and learning new classes. Increasing the value of k progressively biases the training in favor of previously learned classes, thereby allowing fine-grained control over the retention of old class margins. While this strategy effectively reduces negative flips, it may also bias the model against new classes, leading to potential underfitting. To address this issue, we propose a double-source focal distillation mechanism in the following section.

Double-source Focal Distillation Training

When trained with the loss in Equation 2 with a high value of k, the model ϕ tends to preserve the decision margins of old classes but underestimates those of new classes due to the induced positive bias, resulting in significantly smaller decision margins for the latter and a higher error rate. In contrast, training a new independent model from scratch with standard cross-entropy (i.e., without the logit shift) results in more balanced margins and accuracy between old and new classes, but at the cost of increased negative flips.

To address this trade-off, we propose a double-source focal distillation strategy (see Figure 2). This approach transfers positive knowledge from both the old model ϕ_{old} and a new reference model ϕ_{new}, trained on all categories with standard cross-entropy loss, which provides reliable guidance also for the newly added classes. By combining both sources, the final model maintains robustness on old classes while improving classification on new ones, even when sub-

Method	CIFAR100				ImageNet1k			
	ER$_\downarrow$(%) on $\mathcal{Y}^{\text{old}}$	ER$_\downarrow$(%) on $\mathcal{Y}^{\text{all}}$	NFR$_\downarrow$(%)	Rel-NFR$_\downarrow$(%)	ER$_\downarrow$(%) on $\mathcal{Y}^{\text{old}}$	ER$_\downarrow$(%) on $\mathcal{Y}^{\text{all}}$	NFR$_\downarrow$(%)	Rel-NFR$_\downarrow$(%)
Old model	33.27	-	-	-	25.55	-	-	-
No treatment	37.26	36.33	12.20	49.36	22.32	26.38	6.14	36.92
BCT (Shen et al. 2020)	39.08	37.34	13.58	52.38	21.32	25.34	5.71	35.98
PCT-Naive (Yan et al. 2021)	37.38	37.96	12.12	48.88	21.06	26.25	5.30	33.78
PCT-KL (Yan et al. 2021)	36.54	36.98	10.22	42.16	21.78	26.03	5.21	32.14
PCT-LM (Yan et al. 2021)	39.20	41.02	10.50	40.38	24.16	29.38	5.15	28.62
ELODI (Zhao et al. 2024)	35.24	36.62	<u>9.10</u>	<u>39.45</u>	20.42	**24.36**	4.27	28.07
ELODI$_{TopK}$ (Zhao et al. 2024)	35.12	**35.15**	10.14	43.52	22.08	26.16	5.83	35.46
MPT-KL (Ours)	**33.02**	<u>36.07</u>	**8.62**	**39.35**	**19.23**	<u>24.95</u>	<u>3.79</u>	<u>26.48</u>
MPT-LM (Ours)	<u>35.10</u>	37.28	9.42	40.46	<u>19.84</u>	25.50	**3.49**	**23.64**

Table 2: Comparison of different methods on CIFAR100 and ImageNet1k when also the architecture of the new model is updated from a ResNet-18 to a ResNet-50. We report the error rate on old classes (ER on $\mathcal{Y}^{\text{old}}$), the overall error rate (ER on $\mathcal{Y}^{\text{all}}$), the Negative Flip Rate (NFR), and the Relative Negative Flip Rate (Rel-NFR).

jected to an induced positive bias. The proposed final training objective is:

$$\min_{\phi} \mathcal{L}_{\Delta} + \lambda \mathcal{L}_{\text{FD}}(\phi, \phi_{\text{old}}) + \lambda \mathcal{L}_{\text{FD}}(\phi, \phi_{\text{new}}) \tag{3}$$

where $\mathcal{L}_{\Delta}$ is the logit-adjusted softmax cross-entropy loss from Equation 2, and $\mathcal{L}_{\text{FD}}$ a focal distillation loss. We adopt the Focal Loss distillation terms as proposed by Yan et al. (2021) to better emphasize instances correctly classified by the reference model and mitigate the risk of suppressing positive flips. The formal definition of $\mathcal{L}_{\text{FD}}$ is given by:

$$\mathcal{L}_{\text{FD}} = \sum_{i=1}^{N} [\alpha + \beta \cdot \mathbf{1}(\hat{y}_i^* = y_i)] \cdot \mathrm{d}(\phi, \phi^*), \tag{4}$$

where α is a base weight for all samples, β is an additional weight for samples correctly predicted by a reference model ϕ^*, and d is the distance between model outputs. In practice, the distance d can be defined as a temperature-scaled Kullback-Leibler (KL) divergence (Hinton, Vinyals, and Dean 2015):

$$\mathrm{d}_{\text{KL}}(\phi_1, \phi_2) = \text{KL}\left(\sigma\left(\frac{\mathbf{z}_1}{\tau}\right), \sigma\left(\frac{\mathbf{z}_2}{\tau}\right)\right) \tag{5}$$

where σ is the softmax, τ is the temperature, and $\mathbf{z} = \phi(\mathbf{x})$ the logit vector for input $\mathbf{x}$. Alternatively, the Euclidean distance applied directly to the logits may be used (Hinton, Vinyals, and Dean 2015; Buciluǎ, Caruana, and Niculescu-Mizil 2006):

$$\mathrm{d}_{\text{LM}}(\phi_1, \phi_2) = \frac{1}{2} ||\mathbf{z}_1 - \mathbf{z}_2||_2^2. \tag{6}$$

In the following, we refer to this approach as Margin Preserving Training (MPT).

Experimental Results

Experimental Setup

Datasets. We evaluate the proposed approach using two standard image classification datasets: CIFAR100 (Krizhevsky 2009) and ImageNet1K (Russakovsky et al. 2015). ImageNet1K contains 1,000 categories with 1,281,167 training images and 50,000 test images. CIFAR100 comprises 100 categories, including 50,000 training and 10,000 test images. Additionally, we employ CIFAR10 (Krizhevsky 2009) to present qualitative results. For data splits, we follow the methodology from (Yan et al. 2021), using 50% of classes ($\mathcal{Y}^{\text{old}}$) for training the old model and all classes ($\mathcal{Y}^{\text{all}}$) for training the updated model.

Implementation Details. For CIFAR100 and CIFAR10, we train a ResNet-18 and a ResNet-50 (He et al. 2016) using SGD (momentum = 0.9) with a base learning rate of 0.1, applying cosine annealing (Loshchilov and Hutter 2017) over 100 epochs with a batch size of 128. For ImageNet1k, we adopt the standard ImageNet1k training recipe provided directly by PyTorch repository. Specifically, we train a ResNet-18 and a ResNet-50 (He et al. 2016) with embedding dimension 128 using SGD (momentum = 0.9) with an initial learning rate of 0.1, decayed by a factor of ten every 30 epochs for a total of 90 epochs. The batch size is set to 128. For our method, we set $\lambda = 0.4$ and $\lambda = 1.0$ when LM of Equation 6 and KL of Equation 5 are used as distillation functions, respectively. Following prior work (Yan et al. 2021; Zhao et al. 2024), we train every model from scratch, as this represents the most challenging setting, resulting in the highest NFR (Yan et al. 2021).

Evaluation Metrics. To measure model accuracy, we report the error rate (ER) computed on the full test set ($\mathcal{Y}^{\text{all}}$). Additionally, we report ER specifically on test samples from the old classes ($\mathcal{Y}^{\text{old}}$), enabling comparison of the accuracy between base and updated models. As minimizing negative prediction flips (Figure 1) is the main goal of our approach, we adopt the Negative Flip Rate (NFR) (Yan et al. 2021), defined as:

$$\text{NFR} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(\hat{y}_i^{\text{new}} \neq y_i \wedge \hat{y}_i^{\text{old}} = y_i), \tag{7}$$

where N is the number of test samples, $\mathbf{1}(\cdot)$ is the indicator function, $\hat{y}_i^{\text{new}}$ and $\hat{y}_i^{\text{old}}$ denote predictions of the new and old models respectively, and y_i is the ground truth label. Since NFR alone does not account for differing model accuracies, we also report the Relative Negative Flip Rate (Rel-NFR) (Yan et al. 2021), computed as:

$$\text{Rel-NFR} = \frac{\text{NFR}}{(1 - \text{ER}_{\text{old}}) \cdot \text{ER}_{\text{new}}}, \tag{8}$$

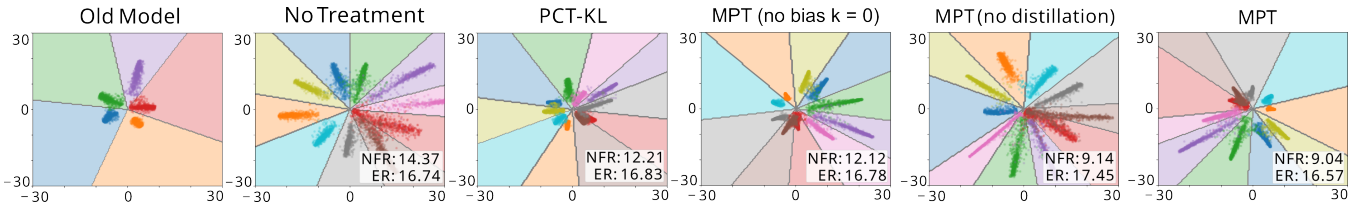

Figure 3: Comparison of embedding spaces for ResNet-18 models (embedding size = 2) on the CIFAR10 test set under various update strategies. Notably, when the model is trained directly on multiple classes ("No Treatment"), the inter-class margins are narrower compared to the original model. Our proposed Margin Preservation Training (MPT) maintains the margins of the original model, thereby significantly reducing the rate of negative flips without compromising the error rate across all classes.

which normalizes negative flips relative to model accuracy differences. The NFR metric measures only cases where previously correct predictions become incorrect. As such, it can decrease even if the overall error on old classes increases slightly. For instance, some examples that were previously misclassified may be correctly classified by the updated model.

Compared Methods. The "Old model" refers to a baseline trained solely on the original dataset (i.e., 50% of the classes). "No Treatment" denotes a setting in which the model is updated using the standard cross-entropy loss, without explicit mechanisms to mitigate negative flips. For comparison, we also include the PCT-Naive approach from (Yan et al. 2021), as well as variants of PCT that employ alternative focal distillation losses: KL divergence (PCT-KL, Equation 5) and Logit Matching (PCT-LM, Equation 6). ELODI (Zhao et al. 2024) is implemented as logit distillation from an ensemble of eight independently trained new models and a single old model. Its TopK variant, which distills only from the Top-K components of the logits, is also included as a baseline (we set K=10 as in Zhao et al. (2024)).

Quantitative Results

We present quantitative comparisons of our proposed MPT methods (MPT-KL and MPT-LM) against existing baselines in Tables 1, where the same architecture is used for both the old and new models, and 2, where the new model adopts a more powerful architecture (ResNet-50) than the original (ResNet-18). The results show that both MPT variants achieve competitive error rates (ER) on old and new classes, indicating effective retention of prior knowledge. Notably, both methods significantly reduce the NFR, highlighting their ability to preserve correct predictions made by the original model. In both experimental settings, MPT-KL and MPT-LM consistently achieve lower NFR values than all baseline methods, notably outperforming ELODI, which employs an ensemble of multiple new models. These findings suggest that preserving the decision margins of the original model plays a key role in mitigating negative flips. The Rel-NFR metric further corroborates this observation: MPT-KL and MPT-LM consistently outperform all baselines, even when controlling for overall accuracy. This indicates that MPT maintains prior knowledge without simply sacrificing accuracy on new classes. Overall, these results clearly demonstrate the effectiveness of the MPT approach

Bias	d	ER($\mathcal{Y}^{\text{old}}$)	ER($\mathcal{Y}^{\text{all}}$)	NFR	Rel-NFR
×	×	28.40	32.88	9.16	43.31
×	LM	27.98	34.70	7.55	36.24
×	KL	27.73	**32.69**	8.09	39.19
✓	×	26.23	34.13	7.71	39.49
✓	KL	**24.68**	<u>32.76</u>	<u>6.09</u>	<u>33.15</u>
✓	LM	<u>25.46</u>	34.12	**6.02**	**31.76**

Table 3: Ablation study of design choices of MPT on ImageNet1k. We report the error rate on old classes (ER($\mathcal{Y}^{\text{old}}$)), the overall error rate (ER($\mathcal{Y}^{\text{all}}$)), the Negative Flip Rate (NFR), and the Relative Negative Flip Rate (Rel-NFR). Variants remove margin-based regularization or distillation. The last two rows report performance of our method, i.e., MPT-KL and MPT-LM.

in reducing negative flips while preserving performance on previously learned classes.

To further validate our approach on modern architectures, we conduct experiments by fine-tuning only the classifier layer of a ViT-B/32 model pretrained on ImageNet-1k, using the CIFAR100 dataset. As shown in Table 4, MPT achieves the best results in terms of error rate for both old and new classes, as well as the lowest value of NFR, thereby validating the effectiveness of our approach.

Qualitative Results

To complement the quantitative analysis presented in previus section, we provide qualitative results obtained from training a ResNet-18 model with a two-dimensional feature space on CIFAR10 using different methods. A two-dimensional feature space enables direct visualization of the learned representations without the need for dimensionality reduction techniques such as t-SNE (Van der Maaten and Hinton 2008) or UMAP (McInnes, Healy, and Melville 2018). Figure 3 presents the resulting representations across six different settings: (1) "Old Model", (2) "No Treatment", (3) PCT-KL, (4) MPT (no bias, $k = 0$), (5) MPT (no distillation, $\lambda = 0$ in Equation 3), and (6) MPT. In each subplot, data points are color-coded according to their true class labels.

In the "Old Model" each class forms a compact cluster centered in its angular sector and is well-separated from adjacent decision boundaries. In "No Treatment" these clusters drift outward and increasingly intersect neighboring sectors:

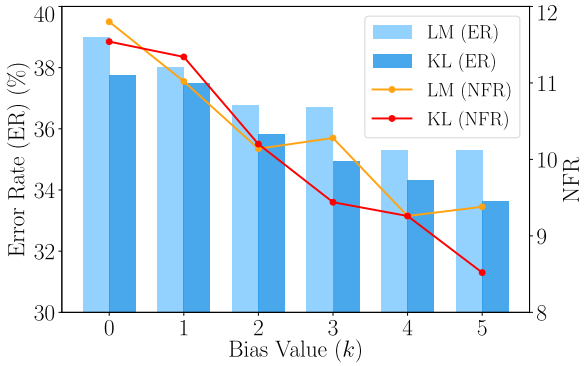

Figure 4: Effect of different margin value k of Equation 1 on CIFAR100 for both LM and KL distances in the focal distillation objectives. Bars show the error rate (ER), and lines represent the negative flip rate (NFR).

the green and blue points cross their boundaries, as do the red and newly introduced brown classes. This indicates substantial margin shrinkage and a tighter geometric arrangement relative to the old model, consistent with the findings of (Wu et al. 2022). The PCT-KL method attempts to replicate the feature distribution of the Old Model, resulting in clusters that are more concentrated near the origin compared to the "No Treatment" scenario. However, not all old class margins are well preserved: some red class points are within the brown class region, indicating a uniform reduction of margins for that class. MPT (no bias, $k = 0$) produces effects similar to those of PCT-KL. By leveraging knowledge from the new model, it prevents new class clusters from collapsing toward the center or into old class regions, resulting in a slightly lower error rate. MPT (no distillation) clearly preserves the margins of the old classes, as all old class points remain within their respective regions. However, this approach results in a high error rate for the new classes, as they are mapped into regions corresponding to the old classes, thereby yielding large decision margins for the new classes. Our proposed approach, MPT, achieves the best results. It provides superior margin preservation for old classes and a lower error rate for new classes, with each point correctly residing in its respective class region. Figure 3 shows the emergence of interfering features with the introduction of new categories, as discussed by Tirer, Huang, and Niles-Weed (2023). For instance, the new brown class (Dog) leads to interference in the learning of the old red class (Cat), as they are semantically similar.

Ablation Studies

Design Choices of MPT. To better understand the contributions of each component in our method, we analize the effects of margin-based regularization and double-source focal distillation. Table 3 shows the impact of removing margin-based regularization ($k = 0$ in Equation 1) or double-source focal distillation by setting $\lambda = 0$ in Equation 3. The results show that both margin-based regularization and focal distillation are essential for minimizing negative flips. Removing either component increases the Negative Flip Rate (NFR) and Relative Negative Flip Rate (Rel-NFR), highlighting their complementary effects. Margin-based regularization alone improves performance on old classes but reduces accuracy on new classes, while distillation alone increases negative flips. Our proposed methods, MPT-KL and MPT-LM, which combine both components, consistently achieve superior performance compared to variants using only one. This confirms that integrating margin preservation and focal distillation is crucial for balanced performance, minimizing negative flips, and maintaining high accuracy across both old and new classes.

Method	ER$_\downarrow$(%) on $\mathcal{Y}^{\text{old}}$	ER$_\downarrow$(%) on $\mathcal{Y}^{\text{all}}$	NFR$_\downarrow$(%)	Rel-NFR$_\downarrow$(%)
Old model	13.28	-	-	-
No treatment	19.44	19.60	7.08	41.86
BCT (Shen et al. 2020)	19.78	19.89	6.64	38.70
PCT-Naive (Yan et al. 2021)	17.40	19.73	5.26	34.85
PCT-KL (Yan et al. 2021)	19.44	19.52	6.88	40.81
PCT-LM (Yan et al. 2021)	19.50	19.59	6.58	38.91
ELODI (Zhao et al. 2024)	17.13	19.54	4.97	30.69
ELODI$_{TopK}$ (Zhao et al. 2024)	17.52	19.57	5.16	31.98
MPT-KL (Ours)	**15.16**	**19.51**	**3.26**	**24.09**
MPT-LM (Ours)	16.22	19.58	3.58	25.45

Table 4: Negative Flip Reduction comparison of a pretrained ViT-B/32 fine-tuned on CIFAR100.

Exploratory Studies of Parameters in MPT. We evaluate the effect of varying the margin parameter k (Equation 1) on the error rate (ER) and Negative Flip Rate (NFR) in CIFAR100, as shown in Figure 4. Owing to computational constraints, these experiments were not repeated on ImageNet-1k. Figure 4 shows that increasing k generally decreases negative flips while maintaining a competitive error rate, with $k = 4$ providing the optimal trade-off for CIFAR100. This trend holds for both KL-based and LM-based distillation objectives. Consequently, we set $k = 4$ for all CIFAR100 experiments. For ImageNet-1k, a smaller margin ($k = 1.5$) is used to better suit the larger dataset.

Conclusions

In this paper, we investigated the problem of negative flips, where updated models misclassify samples previously classified correctly. We addressed this issue from a novel margin-based perspective by explicitly preserving decision margins learned by the old model during incremental updates. Specifically, we proposed Margin Preserving Training (MPT), which applies a logit bias to maintain decision margins for previously learned classes and integrates additional knowledge via a double-source focal distillation strategy. Extensive experiments on CIFAR100 and ImageNet-1k demonstrate that MPT significantly reduces negative flips while maintaining competitive accuracy across all classes.

Limitations. Our method requires training an additional reference model on all classes, modestly increasing computational demands. However, this overhead remains lower than typical ensemble-based methods. MPT's performance depends on tuning of the margin bias, emphasizing the need for empirical optimization across datasets and scenarios.

Acknowledgments

This paper was partially funded by the project "Collaborative Explainable neuro-symbolic AI for Decision Support Assistant", CAI4DSA, CUP B13C23005640006.

References

Abdul Samadh, J.; Gani, M. H.; Hussein, N.; Khattak, M. U.; Naseer, M. M.; Shahbaz Khan, F.; and Khan, S. H. 2024. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. *Advances in Neural Information Processing Systems*, 36.

Amirkhani, A.; Khosravian, A.; Masih-Tehrani, M.; and Kashiani, H. 2021. Robust semantic segmentation with multi-teacher knowledge distillation. *IEEE Access*, 9: 119049–119066.

Anil, R.; Pereyra, G.; Passos, A.; Ormandi, R.; Dahl, G. E.; and Hinton, G. E. 2018. Large scale distributed neural network training through online distillation. In *International Conference on Learning Representations*.

Asadi, N.; Davari, M.; Mudur, S.; Aljundi, R.; and Belilovsky, E. 2023. Prototype-sample relation distillation: towards replay-free continual learning. In *International Conference on Machine Learning*, 1093–1106. PMLR.

Asif, U.; Tang, J.; and Harrer, S. 2020. Ensemble knowledge distillation for learning improved and efficient networks. In *ECAI 2020*, 953–960. IOS Press.

Bahri, D.; and Jiang, H. 2021. Locally adaptive label smoothing for predictive churn. *arXiv preprint arXiv:2102.05140*.

Bansal, G.; Nushi, B.; Kamar, E.; Weld, D. S.; Lasecki, W. S.; and Horvitz, E. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2429–2437.

Bartlett, P. 1996. For valid generalization the size of the weights is more important than the size of the network. *Advances in neural information processing systems*, 9.

Bartlett, P.; Freund, Y.; Lee, W. S.; and Schapire, R. E. 1998. Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics*, 26(5): 1651–1686.

Bartlett, P. L.; Foster, D. J.; and Telgarsky, M. J. 2017. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30.

Beyer, L.; Zhai, X.; Royer, A.; Markeeva, L.; Anil, R.; and Kolesnikov, A. 2022. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10925–10934.

Biondi, N.; Pernici, F.; Bruni, M.; and Del Bimbo, A. 2023a. Cores: Compatible representations via stationarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–16.

Biondi, N.; Pernici, F.; Bruni, M.; Mugnai, D.; and Del Bimbo, A. 2023b. CL2R: Compatible Lifelong Learning Representations. *ACM Transactions on Multimedia Computing, Communications and Applications*, 18(2s): 1–22.

Biondi, N.; Pernici, F.; Ricci, S.; and Del Bimbo, A. 2024. Stationary Representations: Optimally Approximating Compatibility and Implications for Improved Model Replacements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Buciluǎ, C.; Caruana, R.; and Niculescu-Mizil, A. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 535–541.

Cai, D.; Mansimov, E.; Lai, Y.-A.; Su, Y.; Shu, L.; and Zhang, Y. 2022. Measuring and reducing model update regression in structured prediction for NLP. *Advances in Neural Information Processing Systems*, 35: 19384–19397.

Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; and Ma, T. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9650–9660.

Chebotar, Y.; and Waters, A. 2016. Distilling knowledge from ensembles of neural networks for speech recognition. In *Interspeech*, 3439–3443.

Chen, Z.; and Liu, B. 2018. Lifelong machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*.

Delange, M.; Aljundi, R.; Masana, M.; Parisot, S.; Jia, X.; Leonardis, A.; Slabaugh, G.; and Tuytelaars, T. 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4690–4699.

Echterhoff, J. M.; Faghri, F.; Vemulapalli, R.; Hu, T.-Y.; Li, C.-L.; Tuzel, O.; and Pouransari, H. 2024. Muscle: A model update strategy for compatible llm evolution. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 7320–7332.

Elsayed, G.; Krishnan, D.; Mobahi, H.; Regan, K.; and Bengio, S. 2018. Large margin deep networks for classification. *Advances in neural information processing systems*, 31.

Goodfellow, I. J.; Mirza, M.; Xiao, D.; Courville, A.; and Bengio, Y. 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*.

Gunasekar, S.; Zhang, Y.; Aneja, J.; Mendes, C. C. T.; Del Giorno, A.; Gopi, S.; Javaheripi, M.; Kauffmann, P.; de Rosa, G.; Saarikivi, O.; et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. In *NIPS Deep Learning and Representation Learning Workshop*, volume 1050, 9.

Jiang, H.; Narasimhan, H.; Bahri, D.; Cotter, A.; and Rostamizadeh, A. 2021. Churn Reduction via Distillation. In *International Conference on Learning Representations*.

Jiang, J.; Zhou, J.; Wang, P.; Qu, Q.; Mixon, D. G.; You, C.; and Zhu, Z. 2024a. Generalized Neural Collapse for a Large Number of Classes. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 22010–22041. PMLR.

Jiang, Y.; Feng, C.; Zhang, F.; and Bull, D. 2024b. Mtkd: Multi-teacher knowledge distillation for image super-resolution. In *European Conference on Computer Vision*, 364–382. Springer.

Jiang, Y.; Krishnan, D.; Mobahi, H.; and Bengio, S. 2019. Predicting the Generalization Gap in Deep Networks with Margin Distributions. In *International Conference on Learning Representations*.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images. Technical report, Univ. Toronto.

Li, D.; Nguyen, H. L.; and Zhang, H. R. 2023. Identification of Negative Transfers in Multitask Learning Using Surrogate Models. *Transactions on Machine Learning Research*.

Li, Z.; and Hoiem, D. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12): 2935–2947.

Li, Z.; Zhang, M.; Xu, J.; Yao, Y.; Cao, C.; Chen, T.; Ma, X.; and Lü, J. 2023. Lightweight Approaches to DNN Regression Error Reduction: An Uncertainty Alignment Perspective. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, 1187–1199. IEEE.

Liu, J.; Bai, M.; Jiang, N.; Cheng, R.; Li, X.; Wang, Y.; and Yu, D. 2021. Interclass interference suppression in multi-class problems. *Applied Sciences*, 11(1): 450.

Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; and Song, L. 2017. SphereFace: Deep Hypersphere Embedding for Face Recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 6738–6746. IEEE Computer Society.

Liu, Y.; Zhang, W.; and Wang, J. 2020. Adaptive multi-teacher multi-level knowledge distillation. *Neurocomputing*, 415: 106–113.

Lopez-Paz, D.; and Ranzato, M. 2017. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30.

Loshchilov, I.; and Hutter, F. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Malinin, A.; Mlodozeniec, B.; and Gales, M. J. F. 2020. Ensemble Distribution Distillation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

McCloskey, M.; and Cohen, N. J. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, 109–165. Elsevier.

McInnes, L.; Healy, J.; and Melville, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Menon, A. K.; Jayasumana, S.; Rawat, A. S.; Jain, H.; Veit, A.; and Kumar, S. 2021. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*.

Milani Fard, M.; Cormier, Q.; Canini, K.; and Gupta, M. 2016. Launch and iterate: Reducing prediction churn. *Advances in Neural Information Processing Systems*, 29.

Mistretta, M.; Baldrati, A.; Bertini, M.; and Bagdanov, A. D. 2024. Improving zero-shot generalization of learned prompts via unsupervised knowledge distillation. In *European Conference on Computer Vision*, 459–477. Springer.

Neyshabur, B.; Li, Z.; Bhojanapalli, S.; LeCun, Y.; and Srebro, N. 2019. The role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations*.

Papyan, V.; Han, X.; and Donoho, D. 2020. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663.

Parchami-Araghi, A.; Böhle, M.; Rao, S.; and Schiele, B. 2024. Good teachers explain: Explanation-enhanced knowledge distillation. In *European Conference on Computer Vision*, 293–310. Springer.

Pernici, F.; Bruni, M.; Baecchi, C.; and Del Bimbo, A. 2019. Fix your features: Stationary and maximally discriminative embeddings using regular polytope (fixed classifier) networks. *arXiv preprint arXiv:1902.10441*.

Pernici, F.; Bruni, M.; Baecchi, C.; and Del Bimbo, A. 2021. Regular Polytope Networks. *IEEE Transactions on Neural Networks and Learning Systems*.

Pleiss, G.; Zhang, T.; Elenberg, E.; and Weinberger, K. Q. 2020. Identifying mislabeled data using the area under the margin ranking. *Advances in Neural Information Processing Systems*, 33: 17044–17056.

Prabhu, A.; Torr, P.; and Dokania, P. 2020. GDumb: A Simple Approach that Questions Our Progress in Continual Learning. In *ECCV*.

Raffel, C. 2023. Building Machine Learning Models Like Open Source Software. *Commun. ACM*, 66(2): 38–40.

Ren, J.; Yu, C.; Ma, X.; Zhao, H.; Yi, S.; et al. 2020. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33: 4175–4186.

Ricci, S.; Biondi, N.; Pernici, F.; and Bimbo, A. D. 2024. Backward-Compatible Aligned Representations via an Orthogonal Transformation Layer. In *ECCV Workshops (17)*.

Ricci, S.; Biondi, N.; Pernici, F.; Patras, I.; and Bimbo, A. D. 2025. λ-Orthogonality Regularization for Compatible Representation Learning. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*.

Shen, Y.; Xiong, Y.; Xia, W.; and Soatto, S. 2020. Towards Backward-Compatible Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Stanton, S.; Izmailov, P.; Kirichenko, P.; Alemi, A. A.; and Wilson, A. G. 2021. Does knowledge distillation really work? *Advances in neural information processing systems*, 34: 6906–6919.

Tirer, T.; Huang, H.; and Niles-Weed, J. 2023. Perturbation analysis of neural collapse. In *International Conference on Machine Learning*, 34301–34329. PMLR.

Toneva, M.; Sordoni, A.; des Combes, R. T.; Trischler, A.; Bengio, Y.; and Gordon, G. J. 2019. An Empirical Study of Example Forgetting during Deep Neural Network Learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Torralba, A.; Murphy, K. P.; and Freeman, W. T. 2007. Sharing visual features for multiclass and multiview object detection. *IEEE transactions on pattern analysis and machine intelligence*, 29(5): 854–869.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Träuble, F.; Von Kügelgen, J.; Kleindessner, M.; Locatello, F.; Schölkopf, B.; and Gehler, P. 2021. Backward-compatible prediction updates: A probabilistic approach. *Advances in Neural Information Processing Systems*, 34: 116–128.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Wang, T.; Zhou, C.; Sun, Q.; and Zhang, H. 2021. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*.

Weinberger, K. Q.; and Saul, L. K. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2).

Wu, L.; Liu, Z.; Xia, J.; Zang, Z.; Li, S.; and Li, S. Z. 2022. Generalized clustering and multi-manifold learning with geometric structure preservation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 139–147.

Wu, Y.; Chen, Y.; Wang, L.; Ye, Y.; Liu, Z.; Guo, Y.; and Fu, Y. 2019. Large scale incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 374–382.

Xie, Y.; Lai, Y.-A.; Xiong, Y.; Zhang, Y.; and Soatto, S. 2021. Regression bugs are in your model! measuring, reducing and analyzing regressions in nlp model updates. *arXiv preprint arXiv:2105.03048*.

Yadav, P.; Raffel, C.; Muqeeth, M.; Caccia, L.; Liu, H.; Chen, T.; Bansal, M.; Choshen, L.; and Sordoni, A. 2025. A Survey on Model MoErging: Recycling and Routing Among Specialized Experts for Collaborative Learning. *Transactions on Machine Learning Research*.

Yan, S.; Xiong, Y.; Kundu, K.; Yang, S.; Deng, S.; Wang, M.; Xia, W.; and Soatto, S. 2021. Positive-congruent training: Towards regression-free model updates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14299–14308.

Yaras, C.; Wang, P.; Zhu, Z.; Balzano, L.; and Qu, Q. 2022. Neural collapse with normalized features: A geometric analysis over the riemannian manifold. *Advances in neural information processing systems*, 35: 11547–11560.

You, S.; Xu, C.; Xu, C.; and Tao, D. 2017. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 1285–1294.

Yuan, F.; Shou, L.; Pei, J.; Lin, W.; Gong, M.; Fu, Y.; and Jiang, D. 2021. Reinforced multi-teacher selection for knowledge distillation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 14284–14291.

Zenke, F.; Poole, B.; and Ganguli, S. 2017. Continual learning through synaptic intelligence. *ICML*.

Zhang, B.; Ge, Y.; Shen, Y.; Li, Y.; Yuan, C.; XU, X.; Wang, Y.; and Shan, Y. 2021. Hot-Refresh Model Upgrades with Regression-Free Compatible Training in Image Retrieval. In *International Conference on Learning Representations*.

Zhang, L.; Song, J.; Gao, A.; Chen, J.; Bao, C.; and Ma, K. 2019. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3713–3722.

Zhang, W.; Deng, L.; Zhang, L.; and Wu, D. 2022. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*, 10(2): 305–329.

Zhao, Y.; Shen, Y.; Xiong, Y.; Yang, S.; Xia, W.; Tu, Z.; Schiele, B.; and Soatto, S. 2024. Elodi: Ensemble logit difference inhibition for positive-congruent training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhou, Y.; Li, Z.; Shrivastava, A.; Zhao, H.; Torralba, A.; Tian, T.; and Lim, S.-N. 2023. BT^ 2: Backward-compatible Training with Basis Transformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.