We thank the reviewers for insightful comments. We have provided **code** in the supplemental for full reproducibility.

Common Question: The method is for negative transfer rather than catastrophic forgetting.

In transfer learning [23], "transfer" is the ability to apply knowledge learned in previous tasks to new tasks. Due to large domain gap, only part of the pre-learned knowledge is useful for a new task: If the useful part is erased during transfer, it is **catastrophic forgetting**; If the harmful part is preserved during transfer, it is **negative transfer** (Line 31-32).

Hence, catastrophic forgetting and negative transfer constitute a **dilemma** and should be mitigated jointly for optimal performance. This is emphasized by the current Title and Introduction. While catastrophic forgetting has been studied extensively by the community [34, 18], there has no work on mitigating negative transfer in fine-tuning. We propose a novel approach to negative transfer, which is pluggable in the methods for catastrophic forgetting to tackle the dilemma.

R1.1: How is Fig 1 (a)–(d) related to negative transfer? Why the new method works?

Fig 1(a) shows that L^2-SP performs worse than standard fine-tuning L^2. This is a case of *negative transfer* by definition [23]. In **Fig 1(b)** we delve into why negative transfer happens: the *Relative Angles* in the higher layers reveal that the eigenvectors with smaller singular values are **not** transferable. This harmful part causes negative transfer in L^2-SP since it preserves all pre-trained knowledge. Similar results are observed for DELTA [18] (will be added to complete Fig 1).

As justified by [2], with sufficient labeled data, fine-tuning and training from scratch achieve comparably best results—negative transfer does not happen in this case. Hence in **Fig 1(c)–(d)**, we analyze the singular values in this case, and find that the smaller singular values are suppressed more. This hints us that the knowledge conveyed by eigenvectors with smaller singular values are the causes of negative transfer and should be *shrunk*. This well motivates our approach.

R2.1: BSS with continual learning & BSS in text classification with pre-trained word embeddings.

In the table below: **For continual learning**, we evaluate BSS with **EWC** [13] on the permuted MNIST dataset. BSS promotes the target task while slightly hurts the source. **For text classification**, BSS enhances the performance of **BERT** [1], a state-of-the-art NLP pre-trained model. Results on Dev sets are listed, with all hyper-parameters consistent.

Method (continual learning)	task A	task B	Avg	Method (text classification)	MNLI-m	QNLI	MRPC	SST-2
fine-tuning + EWC	96.60	97.42	97.01	$\text{BERT}_{\text{base}}$	84.4	88.4	86.7	92.7
fine-tuning + EWC + BSS	96.46	98.04	97.25	$\text{BERT}_{\text{base}}$ + BSS	85.0	89.6	87.9	93.2

R3.1: Concern on novelty & compare with negative transfer methods in Domain Adaptation.

Orthogonal to catastrophic forgetting, negative transfer is the bottleneck of transfer learning [23] and remains an *open problem* in inductive transfer learning (a.k.a. fine-tuning in the context of deep learning). This work provides the first approach to this important open problem, making a major contribution to this field.

Even in domain adaptation, there lacks in-depth analysis on negative transfer until [32] (CVPR'19). However, domain adaptation and inductive transfer (fine-tuning) are completely different scenarios, detailed in the following table (left). We are the first to address negative transfer in **fine-tuning**, to which domain adaptation methods cannot be applied.

The papers Reviewer #3 lists are important in domain adaptation. We will cite and discuss them in a future version.

Method	labeled source samples	source labels vs. target labels
fine-tuning	unavailable	different
domain adaptation	available	identical

Method	15%	30%	50%	100%
L^2	73.95±0.18	79.43±0.23	81.40±0.21	84.77±0.32
L^2 + re-initialize	70.32±0.32	76.36±0.29	79.98±0.28	83.35±0.33

R3.2: Why not re-initialize all the high-level parameters and train again?

Fig 1 reveals that the eigenvectors with larger singular values in higher layers are transferable. If we re-initialize those parameters, all pre-trained knowledge is discarded and *catastrophic forgetting* happens. Results on Stanford Dogs by re-initializing Layer 4 in ResNet-50 are shown in the above table (right), which are worse than vanilla fine-tuning (L^2).

R3.3: Why use feature regularization instead of parameter regularization?

Parameter regularization has several disadvantages: **(1)** It is hard to decide weights of which layers should be regularized (Line 191-192). In contrast, feature regularization can regularize each layer by taking the advantage of back-propagation. **(2)** The parameters form high-dimensional matrix, whose SVD incurs unacceptable computational cost (Line 207-211). In contrast, we can perform SVD over the feature matrix of each mini-batch, which only adds slightly more computation.

R3.4: Ablation studies of the two parts (catastrophic forgetting & negative transfer).

The ablation studies as requested by the reviewer have already been shown detailedly in Table 2: L^2 denotes the standard fine-tuning; L^2 + BSS is the ablation study for the negative transfer part; L^2-SP / DELTA is the ablation study for the catastrophic forgetting part; And L^2-SP + BSS / DELTA + BSS unifies the two parts to tackle the dilemma in them.

Through above it has been apparent that major questions raised by Reviewer #3 have been answered in the original paper.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

[2] K. He, R. Girshick, and P. Dollár. Rethinking imagenet pre-training. *arXiv preprint arXiv:1811.08883*, 2018.