IHDA+ using Open-sourced Data Augmentation (DA) Policies: We could not test IHDA with advanced DA methods as our initial experiments finished on the last day of the paper submission deadline, and we did not have extra resources to test this. Nevertheless, now we have tested IHDA+ with the learned augmentation policies of AutoAugment (AA) for (a) Wide-ResNet-28-10 on CIFAR (C10) and (b) Resnet-50 on ImageNet. In (a) the test error (%) improved to 1.92 (previously 2.11). In (b) the Top 1/ Top 5 accuracy (%) improved to 81.47 / 96.50 (previously 79.9 / 95.9). These results confirm that if used with advanced DA methods, the IHDA can improve the generalization performance of deep networks even more. Therefore, IHDA can be considered as a complementary approach to the SOTA DA techniques that work in the input space

Comparison with Manifold Mixup (MM) and Adversarial Autoaugment (AAA): We did not compare with MM as it was presented in the literature as a *regularization* technique. As for AAA, we thank the reviewer for pointing it out. Of all the experiments, AAA was better than IHDA & IHDA+ in just two cases (see previous comment). However, based on the new results, IHDA+ with DA policies of AA beats AAA in both settings. We will contrast IHDA with AAA and MM in our paper.

Computational Complexity: IHDA is an iterative method, which starts after the initial training of the model to convergence, where each iteration is a composition of *(a) Generation of augmented data* and *(b) Fine-tuning of the model*. However, the number of iterations is determined by the hyperparameter p, which can be tuned based on practical user constraints. Furthermore, each iteration fine-tunes a smaller version of the model (only proceeding layers are trained) on fewer data points (only points with positive potential are employed) as compared to the initial training. On average, computed over all experiments, IHDA took about 30% of the original training time, which also includes the time spent on tuning hyperparameters. For the sake of comparison, we trained the baseline model for ResNet-110 (without IHDA) for the same extra number of epochs on C10 & C100; the test errors were 6.33 and 28.21, respectively, which are significantly larger than those of IHDA and IHDA+.

Error Plot vs. P: Figure 1 presents test error (%) of ResNet-110 on C10 vs. p for IHDA+.

Novelty: Neither the problem of DA is new, nor is the idea of DA in the feature space, which is the foundation of our method. Nevertheless, our contribution is two-fold: (a) we proposed the *first post-training* DA approach based on generative models that does DA *iteratively* in *difficult regions* of the learned representations to improve the generalization of deep networks. (b) we achieved *better results* than SOTA DA approaches on public benchmarks.

Distance Function: We tried cosine similarity (CS), Euclidean, and Manhattan distances. All gave similar results, but CS's results (reported in the paper) were slightly better.

Preserving Semantics of the Augmented Representations: Although we might think that it is important to preserve the semantics of augmented representations , recent works [Ref:20 from the paper] have shown that DA provides better results if semantic transformations are allowed. In our work, we achieve this through β and ϵ within a generative process.

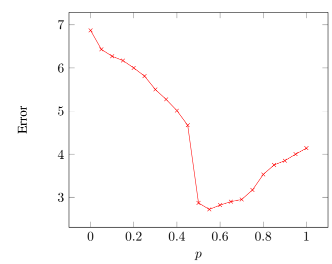

Figure 1: Test error (%) of ResNet-110 vs. p on C10

Combination of Good DA and Self-distillation for a Fair Comparison: We agree that existing DA approaches train the model once; however, most of them do a fair amount of work before that. Nevertheless, we will certainly try to implement their advice and perform a comparison, but it would be extremely helpful if the reviewer explained their idea in more detail.

How Many Examples were Selected in O: As already mentioned on L. 156, we used every example in the set to generate new data points, since every example's potential is positive.

Hyperparameters (HPs): We will mention the values of all HPs in the supplimentary material as best as we can.

Results on ImageNet: The results on ImageNet are reported on a set that is different from the validation set, which was used to tune the hyperparameters. We will clarify this better in the paper.

Initial Accuracy: We have checked our implementation and found that the "Baseline" column represents the initial accuracy of IHDA+. We will also add to the paper the initial accuracy of IHDA.

Beta: Each generated sample has a different β. We tried both with and without β, and empirically found the former to work better. We believe that β provides more powerful semantic transformations in the learned representations.

Others: In the ablation study, P_{φ}^{M} and Random Selection both had $p = 0.55$. We will (a) include the error bounds for as many measurements as possible, (b) expand on differences with [7] in the Related Work, (c) get rid of all the typos.