Use All The Labels: A Hierarchical Multi-Label Contrastive Learning Framework: Supplemental Material

A. Details of Seen and Unseen Data

In the DeepFashion dataset, there are 25,900 training images, 12,612 validation images and 14,218 test images, where we use query images as test images in the task of category classification. We first train our model on seen categories, and finetune the classifier on unseen categories for the task of category classification. For the task of image retrieval, we apply the features from the header to calculate the feature distances between a query image and gallery images. Note that there is no overlap in categories between seen and unseen data, and there is no overlap in images in train and test sets.

ModelNet40 is a synthetic dataset of 3,183 CAD models from 40 object classes, and we split the data into 22 seen and 18 unseen categories. In the seen categories, the numbers of training, validation, and test images are 16,896, 4224, and 5,280, while in the unseen categories, the numbers of them are 13,662, 3,414, and 4,320. For the image retrieval task, the gallery dataset has 11,221 images and the query has 6,017. Since there is no official retrieval split, we split it similar to the validation/test ratio in DeepFashion In-Shop.

B. Qualitative assessment of Image Retrieval on Unseen data

In Figure 1, we show a visualization of the retrieved top-5 images by different algorithms. The top row has 3 query images, and the other rows show their corresponding retrieved top-5 results. The green bounding boxes represent correct retrieved images, and the blue bounding boxes represent wrongly retrieved images but in the correct categories. Recall that the image retreival task is setup at the product level in the hierarchy, which is higher in depth in comparison to the category level. In Figure 1(a), the top-2 retrieved images of the three proposed algorithms are both correct. Although SimCLR and the cross entropy loss do not retrieve correct images, most retrieved images obtain correct categories. In the more challenging example in Figure 1(b), the retrieved images of SimCLR have the best number of correct categories (4), but the corresponding product IDs are not correct among all the 5 retrieved images. In contrast, the proposed HiMulCon and the HiMulConE can both retrieve correct images (top-3). Considering the fact that denims are very similar to pants and the fact that some denims look very similar to each other, e.g. the retrieved images from our proposed algorithms look very similar, our proposed losses have a powerful ability to distinguish similar products. The query image in Figure 1(c) is the most challenging image in these three examples, as it has both tees and denims. We can see that only HiMulConE retrieves the correct image, while all other methods, including the two individual losses that we propose, fail to find the correct product ID. Comparing the results of the proposed three losses to the three baselines, we can see that most retrieved images of our algorithms return a tee-denim combination, which is a reasonable context given the query image. We argue that the combined loss HiMulConE leads to the best learning ability among all methods, with the model showing good separability at both the category and sub-category levels.

C. Ablation Study

C.1. Sampling Strategy

The sampling approach described in main paper tries to sample at least one positive pair from each level in the hierarchy. This strategy ensures that each level has postive pairs, and it also guarantees hard negative mining as the pairs from lower levels can be considered to be hard samples for higher levels. This strategy becomes more relevant with an unbalanced tree structure, as random sampling from a skewed tree structure can lead to the network overfitting to sub-trees with higher image density. For instance, the ratio of image count in the largest and the smallest categories in Deep Fashion training set is over 30. In a statistical study, we found that the random sampling strategy would result in no positive pairs (other than augmented versions of the same image) in over 20% of batches.

We measure the efficacy of our hierarchical batch sampling strategy by comparing its performance with a completely random strategy and a sampling strategy that only ensures multiple positive pairs at the category level. The experiments were all performed with the DeepFashion dataset, with the HiConE loss. All hyperparameters are kept con-

Figure 1. Retrieval visualizations on DeepFashion In-Shop Dataset. The top row has 3 query images, and the rows below show top-5 results of the six methods. Green bounding boxes represent correctly retrieved results, and the blue bounding boxes represent the correct category but wrong product ID.

Approach	Seen	Unseen
Hierarchical batch sampling	**80.52**	**75.29**
Category level sampling	79.81	72.63
Random Sampling	77.96	71.59

Table 1. Ablation study on effectiveness of the batch sampler

stant throughout this set of experiments. Table 1 shows the results, a completely random sampling approach results in a significant deterioration in category prediction.

C.2. Layer Penalty in HiMulCon

The guiding intuition in designing the penalty term in HiMulCon is that higher level pairs need to be forced closer than lower level pairs in the hierarchy. To that end, we evaluate various functions for $\lambda_l = F(l)$, where the functions have a proportional relationship to the level. The functions can also be replaced with an ordered list of penalty values, which can be be treated as tunable hyperparameters, but we leave that for future work.

We evaluate the performance of category prediction on the unseen data validation set of Deep Fashion and the whole validation set of ImageNet for various $f(l)$, and $exp(\frac{1}{l})$ is the candidate picked for other experiments. Keeping with our intuition, note that all of the functions described in Table 2 have an directly proportional relationship with level l. We also performed sanity check experiments where we evaluated various functions that had a inversely proportional relationship as well, their performance was lower than those seen in the table.

C.3. Layer Penalty in HiMulCon: Sanity check experiments

In addition to the ablation study conducted to evaluate different λ_l candidates listed in Table 2, we also conducted some sanity check experiments to verify correctness of the implementation and to validate our intuitive understanding of the effect of the penalty term on the learned representations. Similar to the generalisability experiments in the main paper, all the results reported in Table 3 are conducted on the validation set of the unseen data split in the DeepFashion InShop dataset [2]. In the ablation study covered in Section C.2, we limited our study to functions that were directly proportional to the level. This corresponded to the intuition that the closer the lowest common ancestor of the

$f(l)$	$exp(l)$	$exp(\frac{1}{\|L\|-l})$	2^l	$2^{\frac{1}{\|L\|-l}}$	$\frac{1}{\|L\|-l}$
DeepFashion	73.12	74.29	73.6	73.8	73.47
ImageNet	77.94	79.14	78.36	78.22	78.15

Table 2. Study of various candidates for λ_l for HiMulCon

image pair is to the leaf node, the higher the pair level is, and the higher the penalty should be. In this study, we present experiments with functions that are inversely proportional to the layer level.

First, we evaluated the Identity function $\lambda_l = (\mathbb{1})$, whose effect would be that all layers would be penalized to the same amount, and all layer pairs would be equivalent positive pairs.This function would reduce the HiMulCon loss to become approximately equal to the SupCon [1] loss, but HiMulConE would still benefit from the hierarchical constraint.

Next, we study a collection of exponential functions that decrease with increasing level. The performance significantly deteriorates with increasing proportional penalty terms, validating the relationship between label structure and the loss penalty term.

D. t-sne visualization

We project the test image embeddings into 2 dimensions through t-sne [3] and visualize the results on Deep Fashion dataset in Figure 2. The three proposed losses have a clear category level separability. Interestingly, the semantically similar categories, like *Pants* and *Denim*, as well as *Cardigans* and *Jacket_Coats* are much closer to each other in the embedding space compared to unrelated categories. Although SimCLR and SupCon have some clusters, this is not correlated with category labels, and there is significant mixing of different categories in the clusters from those approaches.

In Figure 3, we present Modelnet40's t-sne visualizations. Consistent with our findings in quantitative analysis in Table 2 of the main paper, we achieve good separability with our approaches. Both SimClr and SupCon have clear sub-spaces in the representation, but they have a poorer correlation with category labels

E. Hierarchy Constraint Study

We defined the hierarchy constraint as the requirement that the loss between image pairs from a higher level in the hierarchy will never be higher than the loss between pairs from a lower level. The hierarchy constraint is violated if the lower level pair has a lower loss than a higher level. In figure 4, we track the frequency of hierarchy violations during the training process on our unseen split of

Approach	Classification Accuracy
HiMulCon, $\lambda_l = \mathbb{1}$	73.1
HiMulConE, $\lambda_l = \mathbb{1}$	73.8
HiMulCon, $\lambda_l = \frac{1}{l}$)	70.8
HiMulCon, $\lambda_l = exp(\frac{1}{l}))$	70.29
HiMulCon, $\lambda_l = 2^{\frac{1}{l}}$)	69.4
SupCon	73.6
HiMulCon, $\lambda_l = exp(\frac{1}{\|L\|-l})$	74.29
HiMulConE, $\lambda_l = exp(\frac{1}{\|L\|-l})$	**76.07**

Table 3. Study of various candidates for λ_l. The last three rows are the results with the validation set on the approaches used in the main paper.

Loss Function	% of pairs violating constraint
HiMulCon	7.45
HiConE	4.26
HiMulConE	**3.74**
SupCon	9.4
SimClr	14.95
Cross Entropy	27.43

Table 4. Hierarchy violations in the test set. Lower numbers indicate fewer pairs violated the hierarchical constraint.

DeepFashion. This value is tracked before the loss is calculated, and is aggregated across all batches in an epoch. All three losses reduce the frequency of constraint violations, but since HiMulCon does not directly optimize for the hierarchy constraint, it does poorly in comparison to the other two losses. Additionally, the penalty term defined by λ_l in HiMulConE helps reduce the frequency of the violations better than HiMulCon.

Next, we studied the frequency of hierarchy violations on the held out test set. We constructed pairs at random, obtained the lowest common ancestor for each pair, computed the distance between the image pairs, and tracked the frequency of hierarchy constraint violation in the embed-

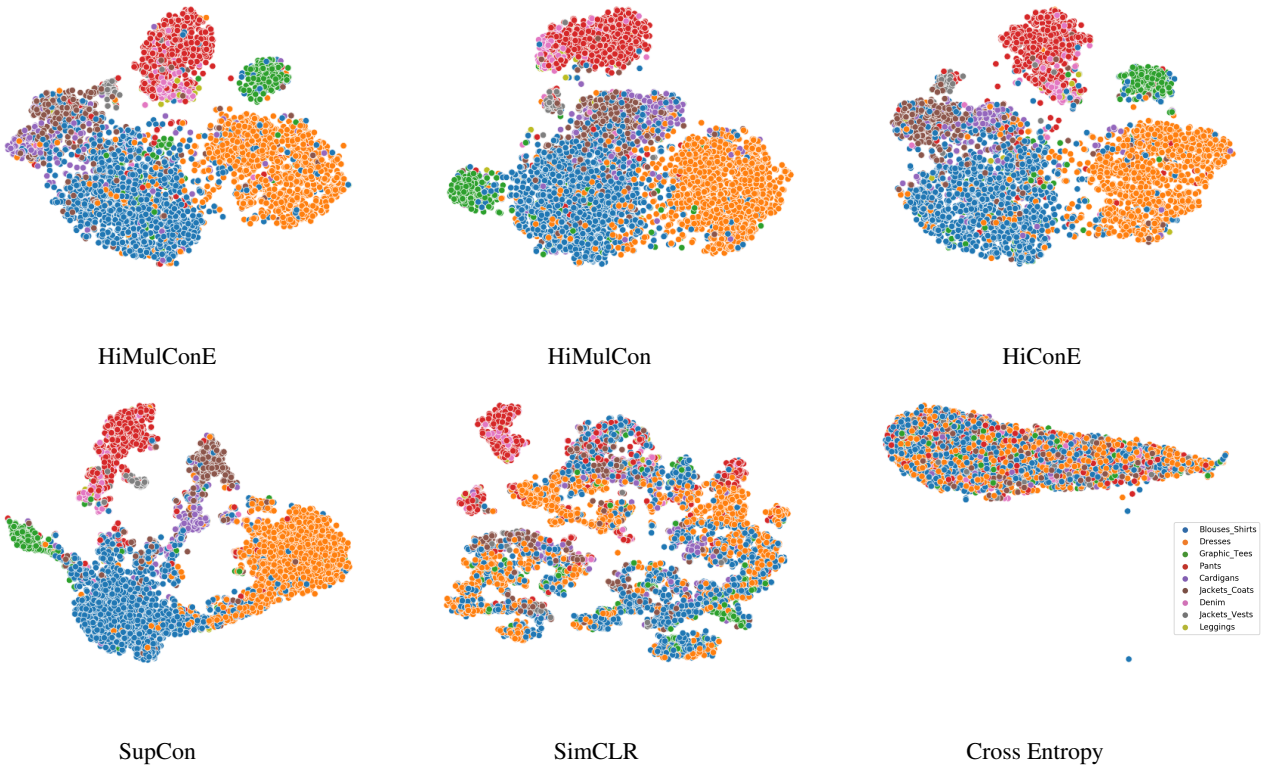

Figure 2. t-sne visualizations of the Deep Fashion dataset

ding space. In the distance based computation, a violation is when a pair from a higher level will be closer than a pair from a lower level. Table 4 presents the hierarchical constraint violation occurrence on the test set. The reduction in hierarchical constraints with convergence, and in the held out validation set, is additional evidence that the representation learning framework presented in this paper preserves the hierarchical label structure in embedding space.

References

[1] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33, 2020. 3

[2] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1096–1104, 2016. 2

[3] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008. 3

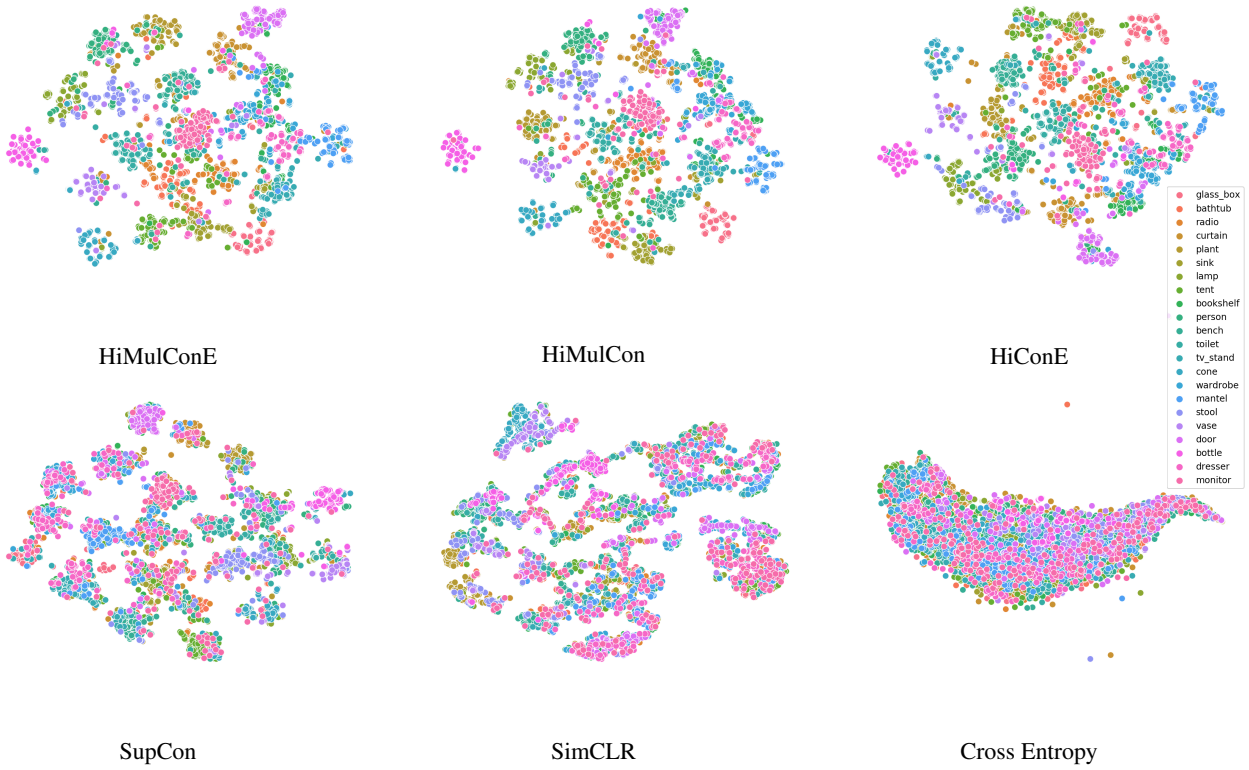

Figure 3. t-sne visualizations of the Modelnet40 dataset. approaches.

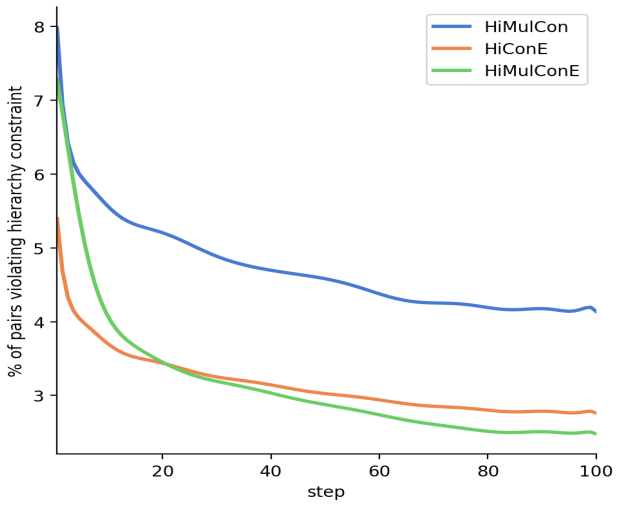

Figure 4. Hierarchical Violations during training