A. Appendix

This supplementary material provides further analyses, additional experimental results, and implementation details that are left out from the main paper due to the space limit.

A.1. Effect of Deploying Semantic Hierarchy

In this section, we investigate the effect of deploying semantic hierarchies by utilizing proxies. In particular, we quantitatively compare our method with SoftTriple loss [6] which can model intra-class variance of data by utilizing multiple proxies per class. Table 2 shows the performance of two proxy-based methods and the number of proxies they used. In SoftTriple loss, 10 proxies are assigned per class, and our method uses 512 hierarchical proxies in addition to proxies for proxy anchor loss.

The result shows that SoftTriple loss lags significantly in performance despite using a much larger amount of proxies compared to our method. The main difference between these methods is how they handle proxies. SoftTriple assigns multiple proxies per class, so their proxies can only represent sub-classes of data and model intra-class variance. On the other hand, proxies in our method can represent sub-classes or super-classes as well as predefined classes, thus allowing more flexibility in modeling more complex relations of data beyond intra-class variance.

A.2. Embedding Space Visualization

We further visualize the embedding space at the beginning of training, such as Epoch 1, 3, 5, 7, and 10. The results of visualization presented in Figure 1 show that the earlier embedding space does not construct a hierarchical structure between the embedding vectors and hierarchical proxies, while this structure is gradually constructed as the training epoch grows. Specifically, as training progresses, the hierarchical proxies have ancestor-descendant relations between sample or other proxies, where these proxies can be regarded as predefined classes, sub-classes, and super-classes. As a consequence, our method can discover the latent semantic hierarchy of training data, which provides rich and granular supervision beyond human-labeled classes.

A.3. More Qualitative Results

We further verify the effect of HIER in a qualitative perspective. To this end, we present the qualitative results of our method, compared to those of a model optimized by only a metric learning loss, $L_{ML}$ in Eq. (9) of the main paper. We take proxy anchor loss [3] as the metric learning loss. Figure 2 presents the qualitative results for the four public benchmark datasets, CUB [9], Cars [4], SOP [7] and In-Shop [5]. These results demonstrate that our method is robust against small inter-class variance (CUB and SOP), viewpoint variation and distinct color (Cars), and large intra-class variations and viewpoint changes (In-Shop) by discovering and deploying a latent semantic hierarchy of data, while the proxy anchor still suffers from those problems.

We note that all the results in the figure are obtained from the fully unseen class samples. It implies that it allows the proposed hierarchical regularizer to discover the latent semantic hierarchy of data with no additional annotation for the hierarchy; our method allows the embedding space to be generalized well.

A.4. Additional Implementation Details

Since each dataset that we utilize to evaluate our method has a different characteristic, we set different hyperparame-
Table 2. Comparison between methods using proxies in terms of performance and the number of proxies on the four benchmark datasets.

<table>
<thead>
<tr>
<th>Methods</th>
<th>CUB #Proxies</th>
<th>R@1</th>
<th>R@2</th>
<th>Cars #Proxies</th>
<th>R@1</th>
<th>R@2</th>
<th>SOP #Proxies</th>
<th>R@1</th>
<th>R@2</th>
<th>R@10</th>
<th>In-Shop #Proxies</th>
<th>R@1</th>
<th>R@2</th>
<th>R@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>SoftTriple [6]</td>
<td>1,000</td>
<td>72.7</td>
<td>82.7</td>
<td>980</td>
<td>83.2</td>
<td>91.3</td>
<td>113,180</td>
<td>80.9</td>
<td>91.3</td>
<td></td>
<td>39,970</td>
<td>88.5</td>
<td>97.3</td>
<td></td>
</tr>
<tr>
<td>PA + HIER (ours)</td>
<td>612</td>
<td>75.2</td>
<td>84.2</td>
<td>610</td>
<td>85.1</td>
<td>91.2</td>
<td>11,830</td>
<td>82.5</td>
<td>92.7</td>
<td></td>
<td>4,509</td>
<td>91.0</td>
<td>98.0</td>
<td></td>
</tr>
</tbody>
</table>

In these experiments, the backbone network is initialized by weights of DeiT [8].

Figure 1. UMAP visualizations of our embedding space learned on the train split of the Cars dataset at different epochs. Pink points indicate hierarchical proxies, and other colors represent distinct classes. The gray line indicates the ancestor-descendant relation between the hierarchical proxy and data points.

References
Figure 2. Qualitative results of ours and proxy anchor on the four public benchmark datasets, CUB (a), Cars (b), SOP (c), and In-Shop (d). Queries and the top 4 retrieval results of our method are presented. The true and false matches are colored in green and red, respectively.