# Supplementary materials

**Anonymous Author(s)**
Affiliation
Address
email

# 1 Appendix

## 1.1 Experimental Configuration

We follow the standard practice of data augmentation in DML [3]: Image are center cropped to be $256 \times 256$ and resized as $288 \times 288$ thereafter. All the models are trained for 40 epochs with batch size 32. We set the learning rate for embedding network as $4e-3$ and learning rate for proxies as $4e2$. We use the Adam optimizer in the model training with no weight decay or learning rate decay. All scripts are written in PyTorch, and run in 2x NVIDIA Titan V GPU.

## 1.2 Performance Comparison of state-of-the-art DML approaches

We show the performance comparison of different DML approaches in Figure 1, among which the top-3 DML papers are all using proxy-based losses.


Figure 1: Recall@1 on CUB200 with different DML approaches

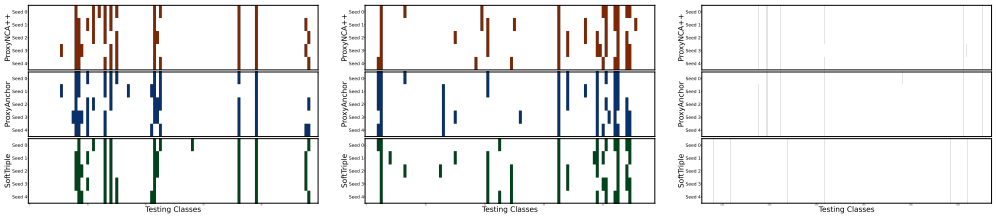
## 1.3 Empirical Study on 3 DML approaches

We start this research with an experiment by running 3 comparable leading DML approaches (i.e., SoftTriple [2], Proxy-NCA++ [3], and Proxy-Anchor [1]) on the popular benchmarks as CUB200, CARS196, and InShop. We observe that there is high overlapping ratio on top-10 wrong testing classes between different trails of training and between different approaches(See Figure 2a, 2b, 2c). We show the testing performance in Table 1.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

| Method \ Dataset | CUB200 | CARS196 | InShop |
|---|---|---|---|
| ProxyNCA++ | 69.04 ± 0.55 | 86.59 ± 0.19 | 86.19 ± 0.19 |
| ProxyAnchor | 68.61 ± 0.97 | 88.76 ± 0.36 | 87.23 ± 0.58 |
| SoftTriple | 67.94 ± 0.53 | 86.47 ± 0.22 | 85.73 ± 0.28 |

Table 1: Recall@1 Performance for 3 approaches on 3 datasets




(a) Empirical study on CUB200 (b) Empirical study on CARS196 (c) Empirical study on InShop

Figure 2: Empirical study: Each row represents a training trail with a specific seed. The first 5 rows are 5 runs from Proxy-NCA++, the next 5 rows are ProxyAnchor, last 5 rows are from SoftTriple. Each column represents a testing class, where the top 10 most frequently wrong testing classes are highlighted. We observe the highlighted columns are well aligned between different runs, different approaches, which indicates they are sharing similar generalization errors.

### 1.4 Mislabel Detection on 1% and 5% Mislabelled Dataset

Figure 3 shows the mislabelled detection accuracy on 1% noisy dataset. And Figure 4 shows the mislabelled detection accuracy on 5% noisy dataset.

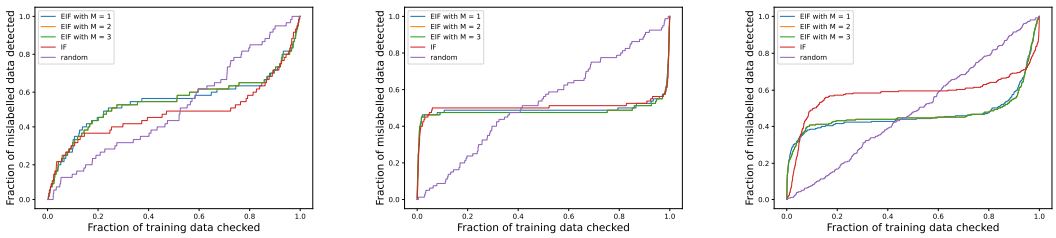



Figure 3: The performance of detecting 1% mislabelled samples on CUB200 (Left), CARS196 (Middle), InShop (Right)

### 1.5 More Examples on Agreeable/Disagreeable Confusion Pair

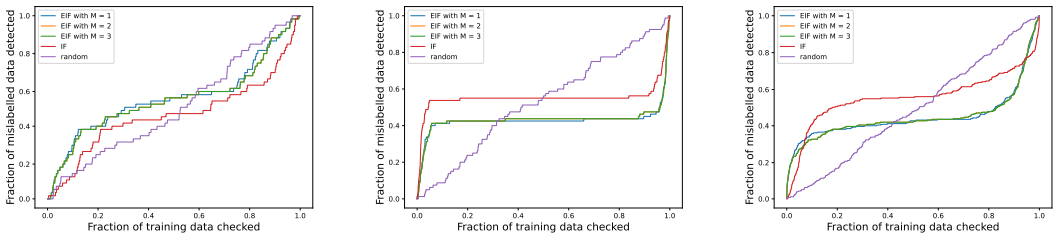
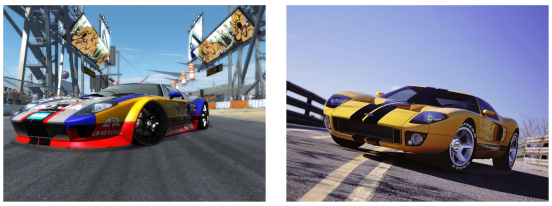



Figure 4: The performance of detecting 5% mislabelled samples on CUB200 (Left), CARS196 (Middle), InShop (Right)

(a) under class 99 (b) under class 98 (c) under class 99

Figure 5: An example of agreeable confusion pair in CARS196 test dataset. Figure 5a and Figure 5b is reported as a confusion pair, and Figure 5a and Figure 5c is labelled as under the same label.



(a) under class 111 (b) under class 111

Figure 6: An example of agreeable mis-similar pair in CARS196 test dataset. Figure 6a and Figure 6b is reported as a mis-similar pair.

(a) under class 99 (b) under class 114 (c) under class 99

Figure 7: An example of disagreeable confusion pair in CARS196 test dataset. Figure 7a and Figure 7b is reported as a confusion pair, and Figure 7a and Figure 7c is labelled as under the same label.

(a) under class 99 (b) under class 99

Figure 8: An example of disagreeable mis-similar pair in CARS196 test dataset. Figure 8a and Figure 8b is reported as a mis-similar pair.

(a) under class 3999

(b) under class 7844

(c) under class 3999

Figure 9: An example of agreeable confusion pair in InShop test dataset. Figure 9a and Figure 9b is reported as a confusion pair, and Figure 9a and Figure 9c is labelled as under the same label.

(a) under class 4018

(b) under class 4018

Figure 10: An example of agreeable mis-similar pair in InShop test dataset. Figure 10a and Figure 10b is reported as a mis-similar pair.

(a) under class 4001

(b) under class 4753

(c) under class 4001

Figure 11: An example of disagreeable confusion pair in InShop test dataset. Figure 11a and Figure 11b is reported as a confusion pair, and Figure 11a and Figure 11c is labelled as under the same label.

(a) under class 4006

(b) under class 4006

Figure 12: An example of disagreeable mis-similar pair in InShop test dataset. Figure 12a and Figure 12b is reported as a mis-similar pair.

## 2 Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (b) Did you describe the limitations of your work? Yes
   (c) Did you discuss any potential negative societal impacts of your work? N/A
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? Yes
   (b) Did you include complete proofs of all theoretical results? Yes
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? Yes
   (c) Did you include any new assets either in the supplemental material or as a URL? Yes
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? N/A
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? N/A
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? N/A

## References


[1] S. Kim, D. Kim, M. Cho, and S. Kwak. Proxy anchor loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3238–3247, 2020.

[2] Q. Qian, L. Shang, B. Sun, J. Hu, H. Li, and R. Jin. Softtriple loss: Deep metric learning without triplet sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6450–6458, 2019.

[3] E. W. Teh, T. DeVries, and G. W. Taylor. Proxynca++: Revisiting and revitalizing proxy neighborhood component analysis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 448–464. Springer, 2020.