Learning Intra-class Multimodal Distributions with Orthonormal Matrices
Supplementary Material

Jumpei Goto, Yohei Nakata, Kiyofumi Abe, Yasunori Ishii
Panasonic Holdings Corporation, Japan
{goto.jumpei, nakata.yohei, abe.kiyo,
ishii.yasunori}@jp.panasonic.com

Takayoshi Yamashita
Chubu University, Japan
takayoshi@isc.chubu.ac.jp

In this supplementary material, we give more visualization examples of the coarse-to-fine experiments (Appendix A), the influence of the label smoothing loss on the proposed method (Appendix B), and the inference time of the proposed model (Appendix C).

A. Visualization examples of coarse-to-fine experiments

In Sec. 4.2 of the main paper, we verified that the proposed method can obtain finer representations than the conventional methods via the coarse-to-fine experiments. We give more visualization examples to stress the interpretability of the features extracted by the proposed method in this section. Figure A1 shows t-SNE visualization results on CIFAR-20 [2] for the "vehicles 1" and "vehicles 2" categories. Specifically, the results reveal that the vanilla technique exhibits overlapping "train" and "bus" features, and although the DNC approach mitigates this propensity, areas with "train," "bus," and "streetcar" features overlapping still exist. Conversely, the proposed method is able to distinguish these three classes; thus, we confirmed that the proposed technique can separate features even in complex feature distributions. Figure A2 shows the nearest neighbor examples of CIFAR-20 training images derived from the query test images. Despite the variations in vehicle type, angle, color of viewing the car, and road surface conditions in the DNC examples, the proposed method demonstrates high interpretability through the similarity in the appearance of the retrieved images. Figures A3 to A6 depict t-SNE visualization outcomes for ImageNet-127 [4] and the nearest neighbor instances of ImageNet-127 training images from the query test images. These visualization results substantiate that the proposed method can extract more interpretable features compared to the conventional methods without relying on fine-grained ground truth labels.

B. Influence of label smoothing loss

As shown in Tab. 3 in the main paper, DNC [7] outperformed the proposed method when utilizing the Swin [3] architecture. In this section, we examined the cause of this performance discrepancy. Since the proposed method incorporates the orthonormal matrices after the backbone, the difference in performance can be attributed to the difference in the process of the classifier between the ResNet and Swin architectures. Thus, we assessed the difference by concentrating on the fact that ResNet [1] employs the cross-entropy loss, whereas Swin uses the label smoothing loss [5].

To investigate the impact of label smoothing loss on the proposed method, we prepared models trained with both the cross-entropy loss and the label smoothing loss. All experimental conditions except for the loss function are the same as the CIFAR-100 [2] experiments. Table A1 shows the top-1/5 accuracy of the proposed model with each loss. Both the top-1 and top-5 accuracies diminish when replacing the cross-entropy loss with the label smoothing loss. Notably, the top-5 accuracy experiences a significant degradation by 1.95 points, suggesting that the proposed method fails to capture inter-class relations using the label smoothing loss. Figure A7 depicts t-SNE visualizations of features extracted by the proposed method using both the cross-entropy loss (a) and the label smoothing loss (b). The label smoothing loss treats all classes other than the ground truth class equally by assigning a constant value to these classes. This property results in more compact intra-class feature distribution as shown in Fig. A7(b) because the features are learned to be mapped away from each other; however, if the distribution of features is too compact, the data multimodality cannot be represented. In fact, we can visually confirm that the distribution of the "lawn-mower" class features in Fig. A7(a) is multimodal, while the distribution in Fig. A7(b) is unimodal. Additionally, it can be observed from Fig. A7 that the label smoothing loss causes the features to lose inter-class relations. For example, while features in the "lawn-mower" and "tractor" categories are plot-

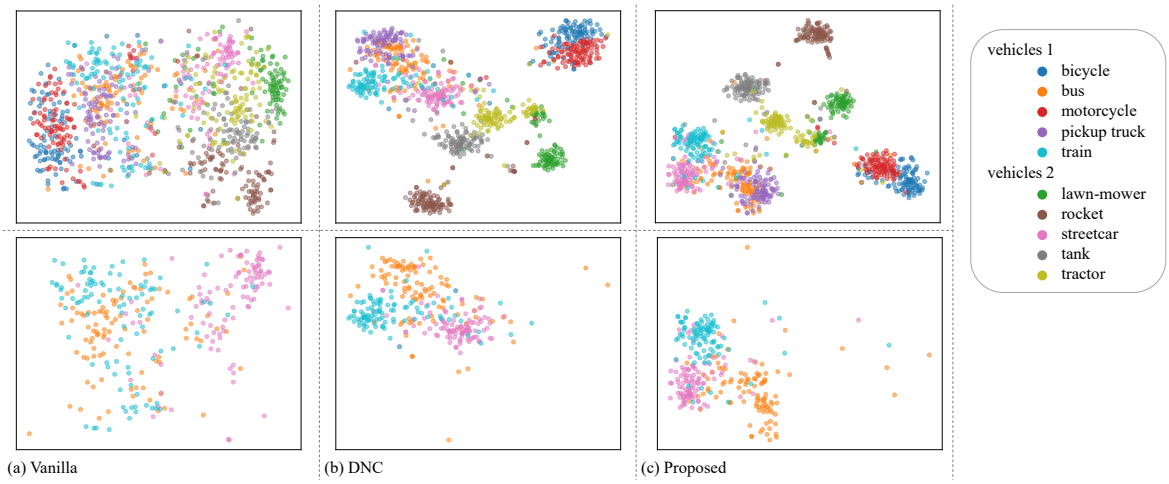

Figure A1. t-SNE [6] visualizations of features extracted from CIFAR-20 [2] images of "vehicles 1" and "vehicles 2" classes (first row). The second row extracts three sub-categories, *i.e.*, "bus," "streetcar," and "train," from the scatters plotted in the first row for better visualization.

Figure A2. Nearest neighbor images from the models trained on CIFAR-20 [2] and evaluated on the fine-grained labels. Correct retrievals are framed with green grids, and incorrect retrievals are framed with orange grids. The query images are from the "bus", "streetcar", and "train" categories.

Table A1. Top-1 and top-5 accuracy (%) on CIFAR-100 [2] validation set. CE and LS represent the cross-entropy loss and the label smoothing loss, respectively.

Method	Loss	top-1	top-5
Proposed	CE	**80.53** ± 0.25	**95.82**
Proposed	LS	79.79 ± 0.25	93.87

Table A2. Inference time (ms) on the ImageNet dataset with a resolution of 224×224.

Method	Backbone	Inference time
Vanilla	ResNet-50	6.604
DNC$_{K=4}$ [7]	ResNet-50	6.950
Proposed	ResNet-50	6.899

ted close together with the cross-entropy loss in Fig. A7(a), features of "lawn-mower" with the label smoothing loss are closer to those of "butterfly" than those of "tractor" in Fig. A7(b), which is against the intuition that a "lawn-mower" is more like a "tractor" than a "butterfly." From these results, we found a direction for future extensions, such as label smoothing that takes inter-class relations into account [8].

C. Inference time

The proposed method incorporates the process of projecting the original backbone features using orthonormal matrices. Thus, we examine the speed impact resulting from the addition of these orthonormal matrices. To evaluate the inference time of the proposed method, we employed the ResNet-50 model and the ImageNet dataset with a resolution of 224x224, measuring the time taken for the model to infer an image. We conducted the time measurement 10,000 times with three methods, which are the vanilla model, DNC, and the proposed model, on a NVIDIA V100 GPU. The mean inference time is reported in Tab. A2. By substituting the d-dimensional weight vectors with the $n \times d$-dimensional weight matrices with $n = 10$ in the classifier, the inference speed of the proposed model becomes only about 4% slower than that of the vanilla model. On the other hand, the inference speed of the proposed model is marginally faster than that of DNC even though the number of the sub-centroids $K = 4$ is smaller than that of the

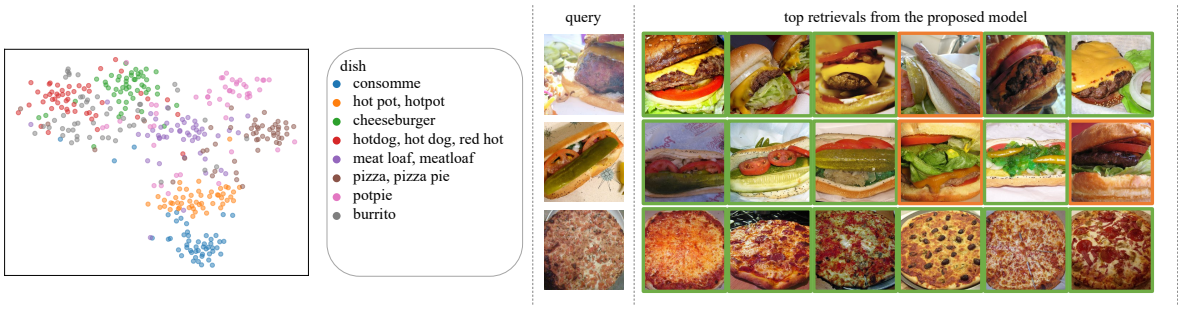

Figure A3. t-SNE [6] visualizations of features that the proposed method extracted from ImageNet-127 [4] images of the coarse-grained label of "dish" (left). Nearest neighbor images from the query images are illustrated on the right. Correct retrievals are framed with green grids, and incorrect retrievals are framed with orange grids. The query images are from the fine-grained categories of "cheeseburger," "hotdog," and "pizza."

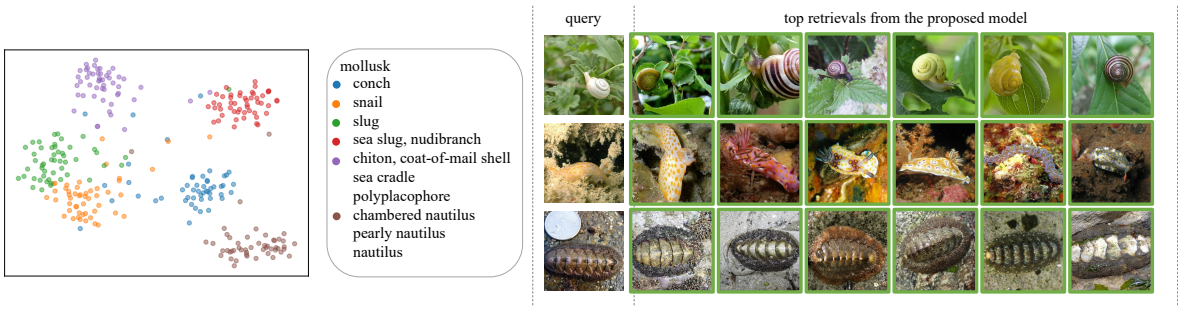

Figure A4. t-SNE [6] visualizations of features that the proposed method extracted from ImageNet-127 [4] images of the coarse-grained label of "mollusk" (left). Nearest neighbor images from the query images are illustrated on the right. Correct retrievals are framed with green grids, and incorrect retrievals are framed with orange grids. The query images are from the fine-grained categories of "snail," "sea slug," and "chiton."

rows of the weight matrices $n = 10$. This can occur because the official implementation of DNC[1] applies the layer normalization to the output logits.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[2] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. 1, 2, 4

[3] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 1

[4] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115:211–252, 2015. 1, 3, 4

[5] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, June 2016. 1

[6] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 2, 3, 4

[7] Wenguan Wang, Cheng Han, Tianfei Zhou, and Dongfang Liu. Visual recognition with deep nearest centroids. In *ICLR*, 2023. 1, 2

[8] Chang-Bin Zhang, Peng-Tao Jiang, Qibin Hou, Yunchao Wei, Qi Han, Zhen Li, and Ming-Ming Cheng. Delving deep into label smoothing. *IEEE Transactions on Image Processing*, 30:5984–5996, 2021. 2

[1]The implementation of DNC is available on https://github.com/ChengHan111/DNC/

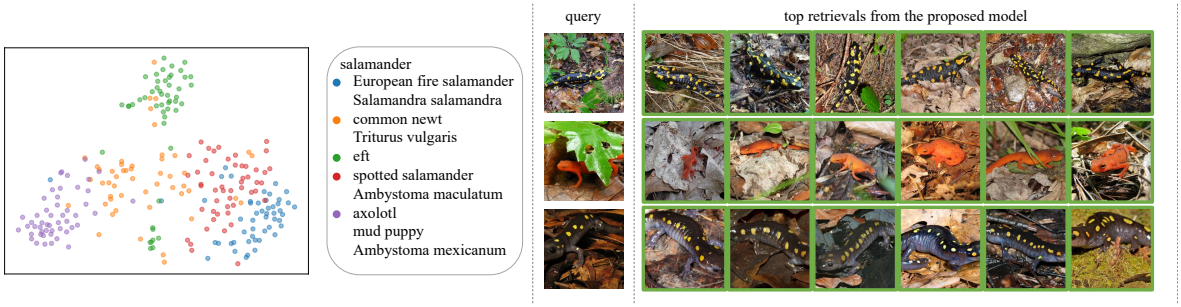

Figure A5. t-SNE [6] visualizations of features that the proposed method extracted from ImageNet-127 [4] images of the coarse-grained label of "salamander" (left). Nearest neighbor images from the query images are illustrated on the right. Correct retrievals are framed with green grids, and incorrect retrievals are framed with orange grids. The query images are from the fine-grained categories of "European fire salamander," "eft," and "spotted salamander."

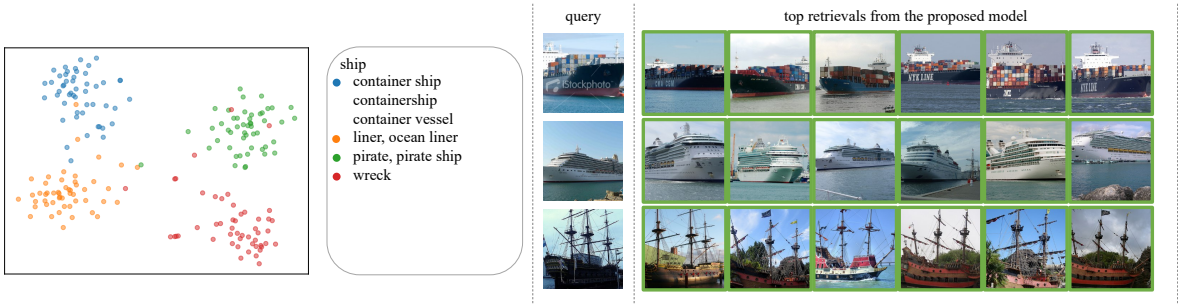

Figure A6. t-SNE [6] visualizations of features that the proposed method extracted from ImageNet-127 [4] images of the coarse-grained label of "ship" (left). Nearest neighbor images from the query images are illustrated on the right. Correct retrievals are framed with green grids, and incorrect retrievals are framed with orange grids. The query images are from the fine-grained categories of "container ship," "liner," and "pirate."

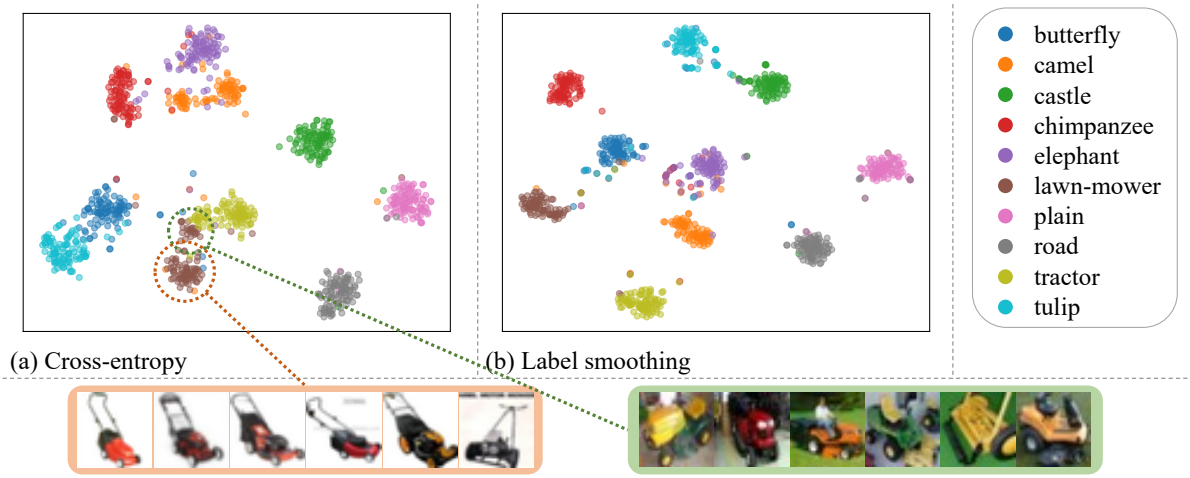

Figure A7. t-SNE [6] visualizations of features that the proposed method extracted from CIFAR-100 [2] images with the cross-entropy loss (a) and the label smoothing loss (b). Ten classes are randomly selected from the CIFAR-100 categories and some images of "lawn-mower" are illustrated at the bottom.