A. Datasets

In this section, we provide more details on the datasets utilized for evaluation. **CIFAR10** [7] contains a total of 60k colorful images in 10 categories. We randomly sample 5k images as the query set and extract 1k samples from each class as the training set, while the rest images are adopted as a retrieval set. **CUB200-2011** [10] is the most common dataset utilized for fine-grained image classification tasks and contains 11,788 examples of 200 subcategories connected with bird species. Following [8], we use the first 100 species for training, while the remaining 100 species are used for testing. **CARS196** [6] contains 16,185 examples of 196 different categories of car models. Following the same setting in [8], we utilize the first 98 species of the car models for training and the rest for testing. **Flickr25k** [3] contains more coarsely grained categories and is more generalized. It contains 25,000 images with some of the 24 labels. We randomly select 50% of the images for training. The remaining images are utilized for testing, i.e., 10% for a query set. **Cars98N** [8] is a real-world noise label benchmark which leverages Pinterest's search engine to obtain a collection of 9,558 examples by utilizing the 98 labels from the CARS196 training dataset as queries.

B. Baselines

We discuss the baseline methods for comparison in details in this section.

- **Fast-AP** [2] & **Smooth-AP** [1]. These two methods are effective variants of the AP-based retrieval approach. Fast-AP optimizes the Average Precision metric based on the interpretation of AP as the area under precision-recall curve while Smooth-AP replaces the indicator function in the AP expression with a sigmoid function to smooth the ranking procedure.

- **Proxy-Anchor** loss [5] is a metric learning loss based on proxies that facilitates rapid and dependable convergence akin to other proxy-based losses. Simultaneously, it capitalizes on the extensive data-to-data relationships during training, resembling the advantages of pair-based losses. In detail, it treats each proxy as an anchor and establishes connections with all data points within the batch. This allows for interactions between samples through the proxy anchor throughout the training process.

- **REL** [11] is a robust early-learning method which can decrease the influence of noisy samples before early stopping when training a neural network. It divides all the parameters into the critical and non-critical ones and then performs different rules to update these two kinds of parameters.

- **HEART** [9] is a noise-resistant hash retrieval method which measures distances between images characterized by their multiple augmented views to choose clean pairs and samples with high confidence. Since HEART is proposed for hash retrieval, we discard the hash layer of the model to fit it into a dense retrieval problem.

- **Jo-SRC** [12] uses JS divergence to measure differences between the given ground truth label distribution and the predicted probability distribution to refine clean samples and leverages contrastive learning techniques to detect OOD/ID samples, then it reassigns labels for OOD/ID noisy samples through a mean-teacher model.

- **T-SINT** [4] is proposed specifically for image retrieval with label noise. It considers all the negative pairs of the samples in a mini-batch to be clean and uses a teacher-based training strategy to recognize false positive pairs and eliminate these false positive pairs from the aggregation process during optimization.

- **PRISM** [8] is a noise-resistant training technique which stores clean features in a memorybank and uses the average similarity of these stored clean features instead of features extracted by the neural network to identify noisy samples.

C. Visulaization Analysis

Challenging Case Study. To compare the retrieval performance of our TITAN and PRISM on a more challenging retrieval benchmark, we return the top-10 results of the query on the more fine-grained CUB dataset, and the results are shown in Figure 1, we can see that the retrieval accuracy on CUB is a bit lower than the retrieval accuracy on coarse-grained benchmarks (e.g., FLICKR25K), and our method still achieves better performance than PRISM, which demonstrates the robustness of our model.

Addtional Qualitative Results. In this part, we plot the Precision-Recall curve and Top-N precision curve of different methods implemented on FLICKR25K to make the qualitative analysis experiments more complete. As shown in Figure 2, our TITAN still achieves the best performance when compared on a larger and more coarse-grained dataset, which demonstrates the generalization and superiority of our method.

D. Supplemental Ablation Analysis

To further explore the effectiveness of more detailed modules in our TITAN, we supplement the additional ablation experimental results with two model variants as follows: (1) **TITAN w/o V** removes the variance of the Gaussian distribution corresponding to the prototype when sampling a feature to Mixing, which means we only use the

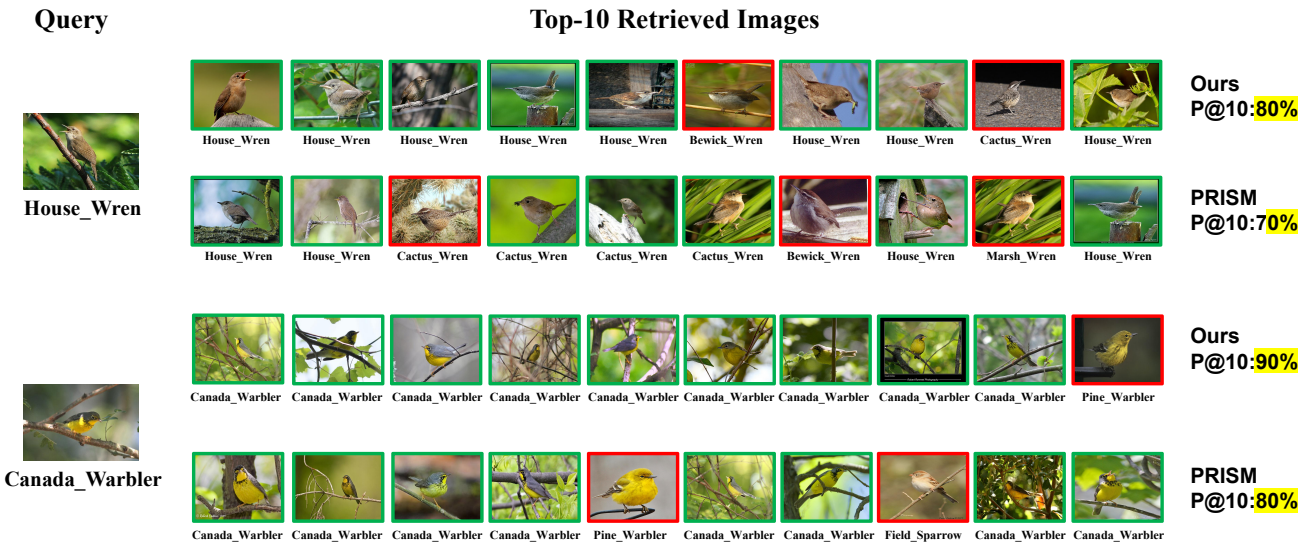

Figure 1. Example of the Top10 returned images with 512-dimensional feature on CUB.

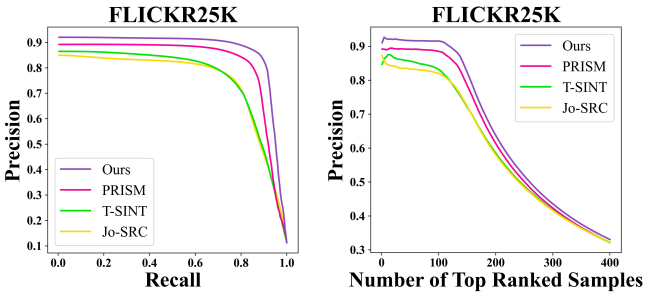

Figure 2. The Precision-Recall curves are plotted in the first column while the Top-N precision curves are in the second column.

Table 1. Addtional ablation results on CIFAR10, CUB, CARS, and FLICKR25K with noise rate being 0.1.

Method	CIFAR10	CUB200	CARS196	FLICKR25K
TITAN w/o V	88.67	18.90	17.17	94.30
TITAN w A	88.01	18.43	16.44	92.34
TITAN(full)	**89.01**	**19.11**	**17.64**	**94.33**

mean vector of the Gaussian distribution. (2) **TITAN w A** replaces sampled virtual feature with randomly augmented feature. The experiment results are shown in Tabel 1, from these results, we can draw some conclusions as follows:

- When the number of training samples is large enough, using a Gaussian distribution to represent the prototype is reasonable. Experimental results show that the MAP@R value decreases slightly when the variance of the Gaussian distribution is removed in the Mixing operation, and the optimization method degenerates to a form closer to the Proxy-based approaches when only the mean vector is used, which implicitly indicates that our method is more robust than the Proxy-based methods.
- When replaces sampled virtual feature with randomly augmented feature, the MAP@R values have a significant decrease, which indicates that our proposed Prototypical Mixing strategy is effective and can alleviate the memory of noisy data to a certain extent.

References

[1] Andrew Brown, Weidi Xie, Vicky Kalogeiton, and Andrew Zisserman. Smooth-ap: Smoothing the path towards large-scale image retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 677–694. Springer, 2020. 1

[2] Fatih Cakir, Kun He, Xide Xia, Brian Kulis, and Stan Sclaroff. Deep metric learning to rank. In *roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1861–1870, 2019. 1

[3] Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, 2008. 1

[4] Sarah Ibrahimi, Arnaud Sors, Rafael Sampaio de Rezende, and Stéphane Clinchant. Learning with label noise for image retrieval by selecting interactions. In *WACV*, pages 2181–2190, 2022. 1

[5] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *CVPR*, pages 3238–3247, 2020. 1

[6] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 1

[7] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[8] Chang Liu, Han Yu, Boyang Li, Zhiqi Shen, Zhanning Gao, Peiran Ren, Xuansong Xie, Lizhen Cui, and Chunyan Miao.

Noise-resistant deep metric learning with ranking-based instance selection. In *CVPR*, pages 6811–6820, 2021. 1

[9] Jinan Sun, Haixin Wang, Xiao Luo, Shikun Zhang, Wei Xiang, Chong Chen, and Xian-Sheng Hua. Heart: Towards effective hash codes under label noise. In *ACMMM*, pages 366–375, 2022. 1

[10] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1

[11] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *ICLR*, 2021. 1

[12] Yazhou Yao, Zeren Sun, Chuanyi Zhang, Fumin Shen, Qi Wu, Jian Zhang, and Zhenmin Tang. Jo-src: A contrastive approach for combating noisy labels. In *CVPR*, 2021. 1