Prototypical Mixing and Retrieval-based Refinement for Label Noise-resistant Image Retrieval

Xinlong Yang[1*], Haixin Wang[1*], Jinan Sun[1], Shikun Zhang[1], Chong Chen[1],
Xian-Sheng Hua[2], Xiao Luo[3†]
[1] Peking University [2] Terminus Group [3] University of California, Los Angeles
{xinlong.yang, wang.hx}@stu.pku.edu.cn, {sjn, zhangsk}@pku.edu.cn, cheung1990@126.com,
huaxiansheng@gmail.com, xiaoluo@cs.ucla.edu

Abstract

Label noise is pervasive in real-world applications, which influences the optimization of neural network models. This paper investigates a realistic but understudied problem of image retrieval under label noise, which could lead to severe overfitting or memorization of noisy samples during optimization. Moreover, identifying noisy samples correctly is still a challenging problem for retrieval models. In this paper, we propose a novel approach called Prototypical Mixing and Retrieval-based Refinement (TITAN) for label noise-resistant image retrieval, which corrects label noise and mitigates the effects of the memorization simultaneously. Specifically, we first characterize numerous prototypes with Gaussian distributions in the hidden space, which would direct the Mixing procedure in providing synthesized samples. These samples are fed into a similarity learning framework with varying emphasis based on the prototypical structure to learn semantics with reduced overfitting. In addition, we retrieve comparable samples for each prototype from simple to complex, which refine noisy samples in an accurate and class-balanced manner. Comprehensive experiments on five benchmark datasets demonstrate the superiority of our proposed TITAN compared with various competing baselines.

1. Introduction

Content-based image retrieval has been an essential research area in computer vision [7], with various applications in search engineering [49, 13] and medical image analysis [19]. Content-based image retrieval can be divided into instance-level retrieval [1, 62] and category-level retrieval [45, 43, 53]. The former focuses on capturing the same instance in different environments. In recent years, category-level retrieval has gained popularity by delivering samples from a massive database with results grouped by the same category as the query [36].

The core of successful category-level retrieval is to map instances to data points in the feature space while preserving similarity connections. The majority of current approaches focus on similarity learning [46, 35], which use pairwise [30] and triplet [16, 50] objectives to maintain semantics in the embedding space. These approaches promote the proximity of semantically similar samples in the embedding space and the separation of semantically different samples. An alternative strategy is to utilize proxies [53, 36, 21, 14]. These methods map each class into the deep feature space to guide semantics learning in a pointwise manner. A few of studies use binary descriptors to improve efficiency [44, 56], which results in a candidate set followed by further refinement for precise image retrieval.

Despite their considerable success, these retrieval systems often presume that the semantic labels in the training set are accurate. In real-world applications, this assumption could be invalidated by the possibility of annotation mistakes and inadequate automated collecting tactics [47, 54]. For instance, web-based sources often include tags and captions, which are usually utilized to provide label information for convenience. These approaches could generate extensive label noise to collected datasets. To address this issue, this work studies a practical topic named image retrieval with label noise. Although there are extensive papers to study the problem of robust learning in classification tasks [29, 12, 25], retrieval models under label noise are still underexplored with unsatisfactory performance in practice.

[*] Equal contribution. [†] Corresponding author.

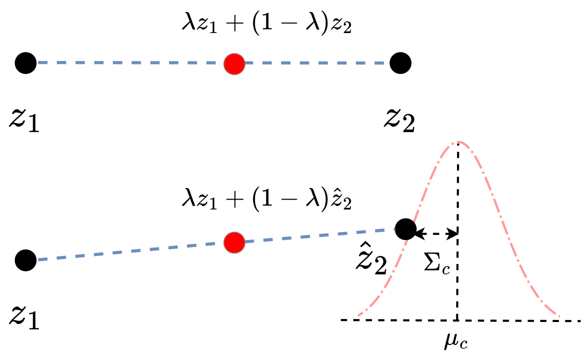
$\lambda z_1 + (1-\lambda) z_2$
z_1
z_2
$\lambda z_1 + (1-\lambda) \hat{z}_2$
$\hat{z}_2$
Σ_c
z_1
μ_c

Figure 1. Two Mixing mechanisms are demonstrated including the standard Mixup (top) and our prototypical Mixing (bottom).

However, developing a label-noise resistant retrieval model is challenging due to two reasons: (1) *Erroneous supervision.* Extensive label noise could result in erroneous supervision information, resulting in a significant decline in retrieval performance. In literature, a few works propose to refine noisy samples for effective image retrieval [28, 18]. For example, PRISM [28] estimates the probability of an example being clean with the help of a memory bank. However, they are prone to result in biased and overconfident refinement due to their point-wise calculation [60]. (2) *Serious Memorization.* Although deep neural networks tend to learn generalized patterns at the beginning stage, the overfitting or memorization of noisy data would compromise the optimization process gradually [57]. As a consequence, a robust optimization objective is anticipated to improve the retrieval performance.

In this study, we propose a novel retrieval method named Prototypical Mixing and Retrieval-based Refinement (TITAN) for label noise-resistant image retrieval. The essence of our TITAN is to measure the distributions of different prototypes in the hidden space. Note that Mixup [58] can generate synthesis samples, which follows the Vicinal Risk Minimization (VRM) [5] principle with high generalization capability. Here, synthetic samples are generated by merging the original data with samples from the respective prototypical distributions, hence avoiding the possible overfitting of noisy samples. Compared with standard Mixup [58], our prototypical Mixing considers more about the hidden structure and does not require mixing different labels. Afterward, we build a similarity learning framework that optimizes the similarity between samples with the same label against using different emphases inferred from prototypical distributions. In addition, to filter potential noisy data, we retrieve comparable samples for each prototype [55]. The identical retrieval number for different prototypes with curriculum learning [2] promises accurate and class-balanced label refinement. Finally, we involve in learning to cluster to enhance the discriminability for effective image retrieval. Comprehensive experiments on a variety of datasets verify the superiority of our TITAN by comparing it to a number of benchmarks. The contribution of this paper can be summarized as follows:

- This paper studies a less-explored but practical problem named label noise-resistant image retrieval and proposes a novel method named TITAN for this problem.
- On the one hand, TITAN measures the distributions of different prototypes and then generates synthesis data to prevent the memorization of noisy samples. On the other hand, TITAN retrieves comparable samples for each prototype from easy to hard, promising accurate and class-balanced label refinement.
- Extensive experiments on five benchmark datasets demonstrate the superiority of our TITAN compared with various baseline methods in different settings.

2. Related Work

2.1. Context-based Image Retrieval

As a basic research topic in computer vision and multimedia communities, context-based image retrieval can be separated into instance-level retrieval [1, 62] and category-level retrieval [45, 43, 53]. Recent years have seen an increasing interest in category-level retrieval, which can be further divided into similarity-based [46, 35] and proxy-based approaches [53, 36, 21, 14]. Similarity-based approaches utilize similarity relationships to generate positive and negative pairs, which are then combined with pairwise [30] and triplet [16, 50] losses to optimize deep neural networks. For instance, GSS [27] tries to distinguish the pairwise similarity in the whole dataset in an unsupervised manner. In contrast, proxy-based approaches map labels into the embedding space before minimizing the distance between deep features and their respective proxies. Hashing-based search engines are also in great demand due to their high efficiency [44, 56]. Nevertheless, these approaches do not often account for label noise in real-world datasets. In contrast, we investigate the issue of label noise-resistant image retrieval and develop a novel framework named TITAN to address it.

2.2. Learning with Label Noise

Learning with label noise has been intensive in a variety of tasks including image segmentation [32], saliency detection [59], and network analysis [9]. A range of works proposes robust losses to potential label noise [11, 10]. For example, RINCE [8] modifies the formula for contrastive learning, which achieves robust performance under noisy augmented views. Another line of this topic is to choose samples with noisy labels and clean them [15, 48, 31, 24, 61, 37, 20]. Early attempts usually put forward a threshold,

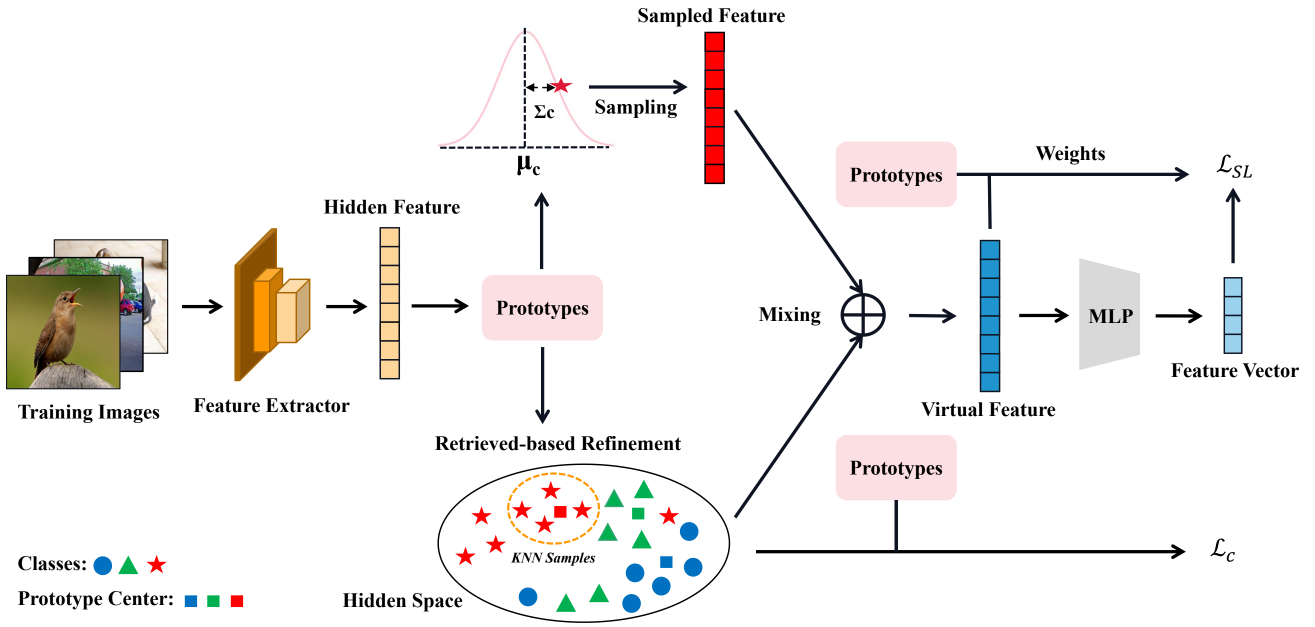

Figure 2. An overview of the proposed TITAN. TITAN first estimates the distributions of prototypes for each class, from which we sample features for Mixing to generate virtual features. Moreover, TITAN retrieves different samples for every prototype to refine noisy samples. The final losses are composed of both similarity learning and clustering losses.

and samples with losses above the threshold are deemed noisy. Co-teaching [15] optimizes two networks concurrently by utilizing small-loss examples from one network to optimize the other network. A few of works attempt to extend robust learning into image retrieval by identifying noisy samples based on similarity [28, 18]. However, these works still neglect the overfitting of noisy samples, and refinement procedures of label noise remain to be biased and overconfident, especially when the amount of noisy data increases. In this paper, we incorporate prototypical Mixing into a similarity learning framework to ease the overfitting of noisy data. To ensure that the labels are assigned correctly and balance class sizes, we also filter noisy samples according to the retrieval process.

3. Methodlogy

3.1. Problem Definition

We start by giving the problem definition. Denote the training set as $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ where $\boldsymbol{x}_i$ is the i-th sample and y_i is its observe labels. n represents the data size. We aim to learn a mapping to project data points into an embedding space where similar image pairs should be mapped into similar deep features and vice versa. In our settings, we assume that the observed labels could be wrong, which increases the difficulty of discriminative feature learning. To evaluate our model, we would retrieve similar examples in the database for the given query sample.

3.2. Overview

Learning effective image representations under label noise requires us to reduce the extreme overfitting of noisy samples and clean up noisy labels in the dataset. Here, we propose a novel method named TITAN for this problem. Following previous works [18, 28], our TITAN modifies a popular classification network (e.g., VGG-F and ResNet50) by adding an MLP layer in place of its head to provide deep features, i.e., $\boldsymbol{h} = H(F(\boldsymbol{x}))$ where $F(\cdot)$ is the feature extractor and $H(\cdot)$ is the MLP head. As illustrated in Figure 2, our model consists of two essential components: (1) *Prototypical Mixing*, which characterizes prototypes using Gaussian distributions and then generates virtual samples by incorporating prototype information to relieve the memorization of noisy samples; (2) *Retrieval-based Refinement*, which retrieves similar samples for every prototype to further refine noisy samples in a balanced way. Then, we elaborate on the details of our TITAN.

3.3. Prototypical Mixing for Robust Representation Learning

The overfitting of neural networks is a major challenge in deep representation learning [57]. As a consequence, we must enhance the generalization capacity of the model. Here, we first characterize the prototypes of various classes using different Gaussian distributions before developing a Mixing technique that blends hidden features with samples derived from these prototypes.

In detail, we characterize the distribution of each class in the prototype as a latent Gaussian distribution $N(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$, and every related hidden feature can be thought of as a sample from the Gaussian distribution [26]. Note that even if these samples could be noisy, the centroid measurement of the distribution is still reliable since all the samples are considered together with potential label noise neutralized. In particular, we measure the centroid and the corresponding covariance matrix based on these samples. In formulation, denote the hidden feature of $\boldsymbol{x}_i$ as $\boldsymbol{z}_i$, we have the estimated mean vector $\boldsymbol{\mu}_c$ and the covariance matrix $\boldsymbol{\Sigma}_c$ of the c-th class in the formulation of:

$$\boldsymbol{\mu}_c = \frac{\sum_{i=1}^{N} \mathbf{1}_{y_i=c} \boldsymbol{z}_i}{\sum_{i=1}^{N} \mathbf{1}_{y_i=c}}, \tag{1}$$

$$\boldsymbol{\Sigma}_c = \frac{\sum_{i=1}^{N} \mathbf{1}_{y_i=c} (\boldsymbol{z}_i - \boldsymbol{\mu}_c) (\boldsymbol{z}_i - \boldsymbol{\mu}_c)^T}{\sum_{i=1}^{N} \mathbf{1}_{y_i=c}}. \tag{2}$$

Then, we generate a virtual sample from the distribution, which would be combined with the original sample. In formulation, we first select the prototype with the maximal probability for each sample, i.e.,

$$\hat{y}_i = \arg\max_c \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_c|}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{z}_i - \boldsymbol{\mu}_c)^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{z}_i - \boldsymbol{\mu}_c)}, \tag{3}$$

where d denotes the hidden dimension. The mixed hidden feature can be written as:

$$\boldsymbol{z}_i^+ = \lambda \boldsymbol{z}_i + (1 - \lambda) \boldsymbol{z}_i', \tag{4}$$

where $\boldsymbol{z}_i'$ is randomly generated from $N(\boldsymbol{\mu}_{\hat{y}_i}, \boldsymbol{\Sigma}_{\hat{y}_i})$, $\boldsymbol{z}_i = F(\boldsymbol{x}_i)$ and $\boldsymbol{x}_i$ denotes the original image sample. λ is a coefficient selected from a Beta distribution following previous works:

$$\lambda \sim Beta(\alpha, \beta) \tag{5}$$

where α and β are two parameters both set to 2 empirically. To simplify the calculation, we adopt a diagonal $\boldsymbol{\Sigma}_c$ for each distribution with the neglection of the covariance.

Advantages of Prototypical Mixing. Similarly to standard Mixup, our prototypical Mixing also generates synthetic samples, extending the Empirical Risk Minimization (ERM) principle [41] to the Vicinal Risk Minimization (VRM) principle [5]. In this way, our TITAN has a strong capacity for generalization and is resistant to the memorization of noisy examples. Moreover, our prototypical Mixing does not need to combine labels to generate synthetic labels, which could be biased since the label space contains distinct relationships. Lastly, our TITAN adequately explores prototypical structures in the hidden space, which can serve as implicit label correction.

Next, we include the mixed samples into a framework for deep representation learning that maintains similarity. Here, we concatenate the hidden feature with the mixed feature into a mini-batch and maximize the similarity between deep features of positive pairs compared with negative pairs. The positives for $\boldsymbol{z}_i^+$ are mini-batch samples sharing the same label, i.e., $\boldsymbol{z}_j^+$ with $y_i = y_j$. In formulation, given the mini-batch $\mathcal{B}$, the index of positives for $\boldsymbol{z}_i^+$ is written as:

$$\Pi(i) = \{j | \boldsymbol{x}_j \in \mathcal{B}, y_i = y_j\}. \tag{6}$$

Following previous works [38, 6], the negatives are all the other samples in a mini-batch, then the similarity learning objective is constructed as:

$$\mathcal{L}_{SL} = -\sum_{\boldsymbol{x}_i \in \mathcal{B}} \frac{1}{|\Pi(i)|} \sum_{j \in \Pi(i)} \log \frac{\exp\left(\boldsymbol{h}_i^+ \cdot \boldsymbol{h}_j^+ / \tau\right)}{\sum_{\boldsymbol{x}_{j'} \in \mathcal{B}} \exp\left(\boldsymbol{h}_i^+ \cdot \boldsymbol{h}_{j'}^+ / \tau\right)} \tag{7}$$

where $\boldsymbol{h}_i^+ = H(\boldsymbol{z}_i^+)$ maps the hidden feature into the target space and τ is a temperature parameter preset to 0.5 as in [38]. Moreover, we also incorporate the probabilities in Eqn. 3 as the confidence scores, which gives more emphasis on samples with high confidence. Consequently, we alleviate the effect of unreliable mixed data and enlarge the range of sample weights. In other words, sample weights are assigned more heavily to reliable samples. Let $\boldsymbol{w}_i = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_{\hat{y}_i}|}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{z}_i - \boldsymbol{\mu}_{\hat{y}_i})^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{z}_i - \boldsymbol{\mu}_{\hat{y}_i})}$ denote the confidence score, and we rewrite Eqn. 7 as:

$$\mathcal{L}_{SL} = -\sum_{\boldsymbol{x}_i \in \mathcal{B}} \frac{\boldsymbol{w}_i}{|\Pi(i)|} \sum_{j \in \Pi(i)} \log \frac{\exp\left(\boldsymbol{h}_i^+ \cdot \boldsymbol{h}_j^+ / \tau\right)}{\sum_{\boldsymbol{x}_{j'} \in \mathcal{B}} \exp\left(\boldsymbol{h}_i^+ \cdot \boldsymbol{h}_{j'}^+ / \tau\right)}. \tag{8}$$

Compared with proxy-based methods [53, 36, 21, 14], our similarity learning-based objective makes use of pairwise labels, which can be more accurate due to the following reason. Even if two labels in a positive pair are both incorrect, there is still a chance that they share the same semantics [51]. Our Mixing strategy further relieves the memorization of noisy samples, producing label noise-resistant and similarity-preserving deep features.

3.4. Retrieval-based Refinement with Curriculum Learning

We need to identify and refine those noisy samples to thoroughly get rid of label noise for better performances [53, 36, 21, 14]. Traditional classification techniques consider samples with small training losses to be clean samples [15], and they often introduce a threshold to distinguish between clean and noisy data. Nevertheless, we are unable to directly obtain losses for various samples in the similarity learning framework. Even worse, a low threshold could exclude too many noisy samples resulting in potential underfitting, while a high threshold may have a limited capacity to filter noisy data. These point-wise methods could also be biased to refine noisy samples into easy classes with small losses [33]. To address the problems, we provide a retrieval-based refinement module, which retrieves comparable samples for each prototype with curriculum learning [2] for precise and balanced refinement.

In particular, we view the centroid of each prototype as a query and retrieve K samples in the dataset based on the Euclidean distance. For these retrieved samples, we regard them as clean if their labels are not consistent with their queries, otherwise refine their labels. In formulation, the refined datasets can be written as below:

$$\begin{aligned}\hat{\mathcal{D}} = \cup_{c=1}^{C} \{(\boldsymbol{x}_i, c)|\boldsymbol{z}_i \in NN(\boldsymbol{\mu}_c)\} \\ \cup \{(\boldsymbol{x}_i, y_i)|\boldsymbol{z}_i \notin \cup_{c=1}^{C} NN(\boldsymbol{\mu}_c)\},\end{aligned} \tag{9}$$

where $NN(\cdot)$ denotes the set of top K nearest samples in the datasets. Here we set the same K for every prototype, which generates an accurate and balanced dataset after refinement. In comparison, clean sample selection based on losses could lean toward selecting clean samples from easy classes. Moreover, we propose to increase the retrieval number K gradually following the principle of curriculum learning, which refines noisy samples from easy to hard. To be specific, we have different K_t at the t-th cycle:

$$K_t = \frac{t}{T} K_{max}, \tag{10}$$

where T denotes the total number of cycles and K_{max} denotes the maximal retrieval number. To increase framework efficiency, we would conduct retrieval-based refinement at the beginning of each cycle.

3.5. Summary

Learn to Cluster. Finally, we enhance the discriminability of hidden features by learning to cluster [40]. This is accomplished by enforcing the accumulation of hidden features around their corresponding centroids. Here, we maximize the similarity between hidden features and their corresponding prototypes compared with the other prototypes by proposing the following objective:

$$\mathcal{L}_C = - \sum_{\boldsymbol{x}_i \in \mathcal{B}} \log \frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{\mu}_{y_i}/\tau)}{\sum_{c=1}^{C} \exp(\boldsymbol{z}_i \cdot \boldsymbol{\mu}_c/\tau)}, \tag{11}$$

Then the final objective is summarized by combining both two losses as follows:

$$\mathcal{L} = \mathcal{L}_{SL} + \eta \mathcal{L}_C, \tag{12}$$

where η is a parameter to balance two losses. In practice, we would first warm up the neural network without conducting Mixing and then perform prototypical Mixing and retrieval-based refinement gradually. We would demonstrate the whole algorithm in Algorithm 1.

Algorithm 1 Training Algorithm of TITAN

```
Require: Dataset $\mathcal{D}$; Iteration number $T$; Max retrieval number $K_{max}$;
Ensure: The composited projector $H(F(\cdot))$;
Warm up the projector;
repeat
    Update the retrieval number $K_t$ using Eqn. 10
    Estimate the distribution for every prototype using Eqn. 1 and Eqn. 2;
    Refine the dataset using Eqn. 9;
    for $k = 1, 2, \cdots, k_{max}$ do
        Sample $\mathcal{B} \subset \hat{\mathcal{D}}$ to construct a mini-batch;
        Generate synthetic samples using Eqn. 4 ;
        Calculate the loss objective by Eqn. 12;
        Update the network parameters by backpropagation;
    end for
until convergence
```

4. Experiment

4.1. Settings

Datasets. In order to evaluate the TITAN, we utilize four widely used datasets with various levels of granularity and size and one real-world noisy dataset. The concrete experimental setting is depicted in the Appendix. **CIFAR10** [23] contains a total of 60k colorful images in 10 categories. **CUB200-2011** [42] is the most common dataset utilized for fine-grained visual classification tasks and contains 11,788 images of 200 subcategories related to bird species. **CARS196** [22] includes 16,185 samples of 196 distinct categories of car models. **FLICKR25K** [17] contains more coarsely grained categories and is more generalized. **Cars98N** [28] is a real-world noisy benchmark, it's built by collecting 9,558 examples for 98 car models from Pinterest. Normally the noisy examples in Cars98N usually consist of the car parts, the interior of the car, or pictures of other car models.

To investigate the noise-resistant ability, we firstly synthesize two kinds of artificial noise labels, Symmetric and Pairflip, added to the training set of the none real-world noise dataset. Following [28], we vary the noise ratio with three degrees: 10%, 20%, and 50% respectively. Symmetric noise [34] simply flips labels between classes uniformly while Pairflip noise [34] similarly assigns the corrupted label to each nearby class.

Baselines. We compare our TITAN with eight baselines that have excellent performances in their corresponding field. Concretely, three of these methods are standard retrieval approaches (i.e., Fast-AP [4], Smooth-AP [3] and Proxy-Anchor [21]), two are robust noise-resistant approaches (i.e., REL [52] and Jo-SRC [54]), one is noise-

Table 1. The MAP@R scores implemented on 512-dimensional feature vectors for retrieval on CIFAR10, CUB, CARS, and FLICKR25K with three degrees of symmetric label noise.

Method	CIFAR10			CUB200			CARS196			FLICKR25K		
	0.1	0.2	0.5	0.1	0.2	0.5	0.1	0.2	0.5	0.1	0.2	0.5
FastAP	85.07	83.40	65.92	13.85	12.81	8.63	10.78	9.28	6.47	88.95	88.59	87.50
SmoothAP	84.48	81.38	70.11	12.71	11.78	9.78	9.70	8.82	6.19	86.37	85.20	82.30
Proxy-Anchor	87.20	78.04	69.23	16.70	15.86	10.11	12.21	11.11	5.33	89.12	88.69	87.79
Jo-SRC	87.58	85.55	72.86	15.97	14.91	12.91	12.74	11.09	7.10	91.01	90.22	89.43
REL	86.81	84.76	71.90	15.78	15.67	13.86	14.02	13.33	7.96	91.59	91.44	90.60
HEART	85.36	83.15	71.65	17.10	15.79	12.00	15.30	13.59	8.64	92.11	91.81	91.01
T-SINT	87.38	86.26	72.87	17.90	17.36	14.34	14.41	13.01	9.89	90.79	90.45	87.65
PRISM	87.65	86.23	73.05	18.12	17.79	15.29	16.99	15.88	10.02	92.42	92.12	91.23
TITAN (Ours)	**89.01**	**88.01**	**75.86**	**19.11**	**18.69**	**16.90**	**17.64**	**16.36**	**11.37**	**94.33**	**93.89**	**93.07**

Table 2. The MAP@R scores implemented on 512-dimensional feature vectors for retrieval on CIFAR10, CUB, CARS, and FLICKR25K with three degrees of pairflip label noise.

Method	CIFAR10			CUB200			CARS196			FLICKR25K		
	0.1	0.2	0.5	0.1	0.2	0.5	0.1	0.2	0.5	0.1	0.2	0.5
FastAP	77.42	76.52	56.27	14.06	12.42	11.06	10.60	8.53	6.36	87.53	87.16	85.22
SmoothAP	76.15	74.55	55.86	12.12	11.42	10.56	10.62	9.55	6.84	85.30	85.18	85.02
Proxy-Anchor	75.55	73.82	57.65	16.08	15.11	12.16	13.52	11.45	8.61	90.19	89.26	88.23
Jo-SRC	78.52	76.24	57.77	15.92	14.55	13.45	14.48	13.43	9.20	91.63	91.60	90.11
REL	75.83	75.12	58.72	15.76	14.78	13.05	12.50	11.72	9.09	91.46	90.78	89.19
HEART	77.82	76.11	58.55	16.26	15.33	13.07	13.58	13.36	8.77	91.76	91.72	90.99
T-SINT	79.19	77.94	55.39	16.51	16.14	14.26	14.45	13.69	9.07	91.62	90.74	89.71
PRISM	80.43	80.32	62.05	17.13	15.83	13.13	15.46	15.15	9.38	92.44	92.27	90.91
TITAN (Ours)	**83.89**	**82.99**	**67.88**	**18.29**	**17.36**	**15.73**	**16.98**	**15.82**	**10.14**	**94.05**	**93.90**	**93.18**

robust hash retrieval method (i.e., HEART [38]), two are noise-robust retrieval methods (i.e., T-SINT [18] and PRISM [28]).

Table 3. The MAP@R scores implemented on 512-dimensional feature vectors for retrieval on Cars98N.

Method	TITAN(Ours)	T-SINT	PRISM	HEART	REL	FastAP
Cars98N	**7.03**	6.53	6.25	5.54	5.38	5.06

Evaluation Criterion. Following [28, 18], we mainly employ Mean Average Precision@R (MAP@R) as our criterion for evaluation. MAP@R metric is a widely used measure of average accuracy for multiple queries and is a common metric for evaluating the quality of retrieval systems. Note that the value of R is not a fixed value for all benchmarks, it is equal to the largest number of samples among all categories in the corresponding retrieved set of the benchmark.

Implementation Details. We implement our method using PyTorch with an NVIDIA 3090 GPU. We adopt mini-batch Adam for our model training. The mini-batch size is set to 64 and the learning rate for our model is fixed at 5×10^{-5}. All the images are resized to 224×224, and the training images have a 50% chance of being randomly flipped horizontally. In all experiments, we use ResNet18 as the feature extractor with the feature dimension being 512, and following [28] we use L2 normalization to make the output of the neural network be on the unit hypersphere for brief distance measurement. In particular, we also choose ResNet18 instead of CLIP as the teacher model when reimplementing T-SINT for a fair comparison.

4.2. Experimental Results

Quantitative Comparison. Table 1 and Table 2 display the MAP@R results on CIFAR10, CARS, CUB, and FLICKR25K under Symmetric/Pairflip label noise specifically with noise rate being 0.1, 0.2, and 0.5. From these experimental results we can make the following observations: (1) The AP-based retrieval methods (i.e., FastAP and SmoothAP) often fail to perform as well as proxy-based retrieval methods in most cases. This may be due to the proxy-based approach's use of the feature centers of each class to protect metric learning from label noise. (2) The noise-resistant classification methods (i.e., Jo-SRC

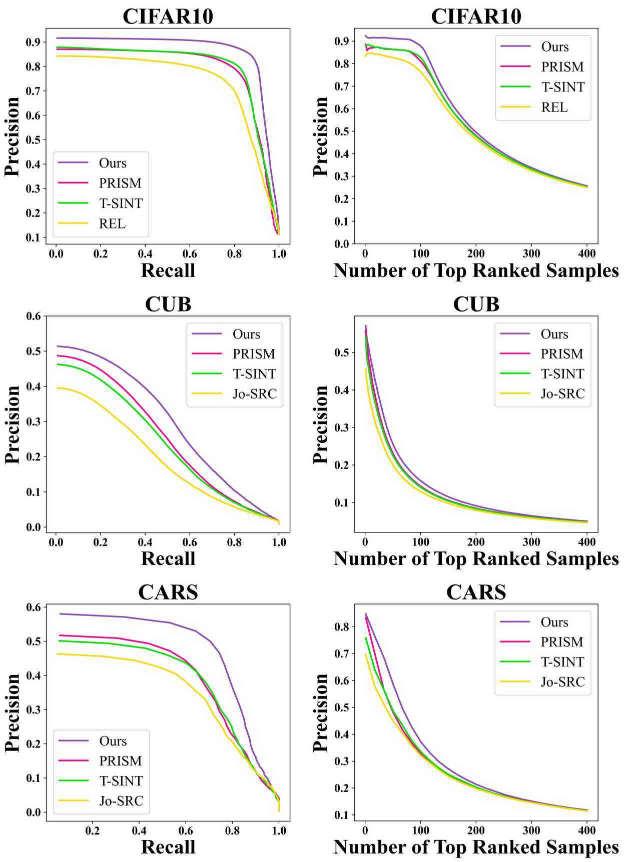

Figure 3. The P-R curves are plotted in the left column and Top-N precision curves are in the right column.

and REL) achieve limited performance increases compared with standard retrieval methods, which illustrates that point-based classification approaches are not suitable for retrieval tasks under label noise and special techniques need to be designed to solve the problem. (3) Our proposed TITAN achieves the best scores on all the benchmarks with different settings. For CUB200, our TITAN obtain an improvement of 5.46%, 5.05%, and 9.52% over the best baseline PRISM under symmetric label noise with noise rate being 0.1, 0.2, 0.5 respectively. For CARS and FLICKR25K, our TITAN outperforms competitive baseline T-SINT by 21.04% and 4.66% in terms of average MAP@R scores under symmetric noise with different noise degrees. And the same tendency can also be observed on CIFAR10 under both noise label settings. The exceptional performance of our TITAN can be attributed to two primary factors. On the one hand, we improve the accuracy of noisy sample refinement by using retrieval-based refinement with curriculum learning instead of point-wise refinement. Our model, on the other hand, is enhanced by prototypical mixing technology that prevents it from memorizing noisy data, which facilitates semantic learning.

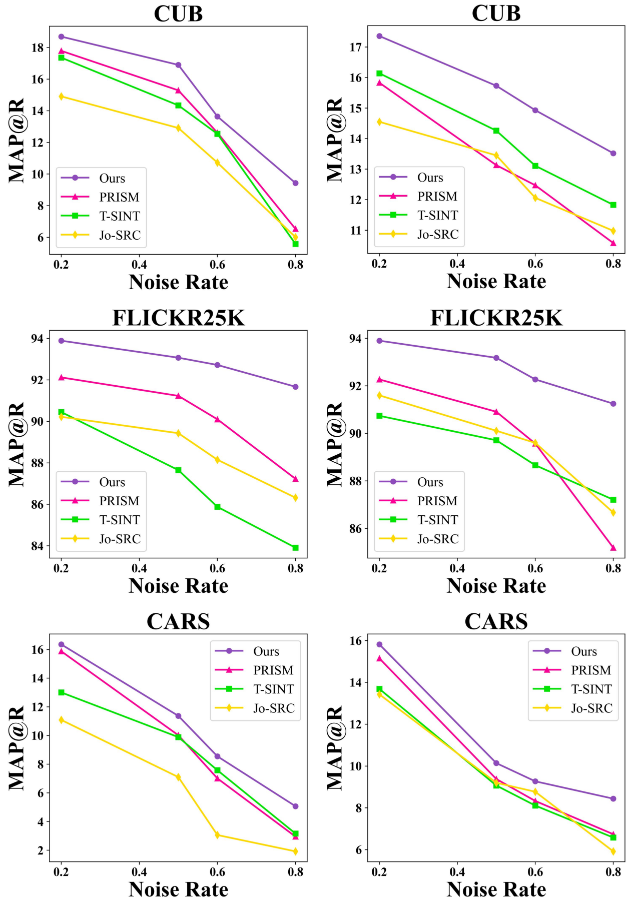

Figure 4. The MAP@R scores w.r.t. different noise rates. Results with Symmetric noise are plotted in the left column while results with Pairflip noise are plotted in the right column.

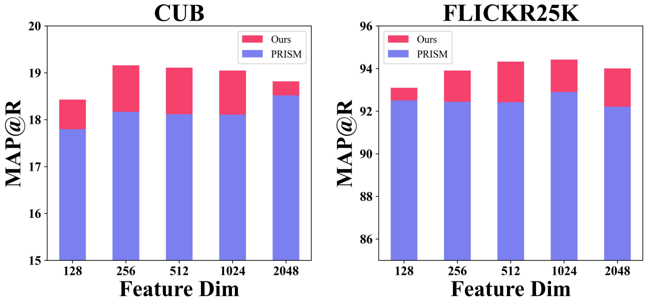

Figure 5. The MAP@R scores of TITAN and PRISM implemented on CUB and Fickr25k with varying hidden dimensions while the noise rate is set to 10%.

Table 3 displays the performance on Cars98N with real-world label noise. From these results, we can observe that TITAN also achieves significantly better performance in real-world settings, which demonstrates the generalizability and superiority of our approach.

Qualitative Comparison. Additionally, we plot the P-R curves and TopN-precision curves of the proposed TITAN and other compared methods with 512-dimensional feature vectors on three benchmarks in Figure 3. In the P-R curve,

Table 4. Ablation results on CIFAR10, CUB, CARS, and FLICKR25K with noise rate being 0.1.

Method	CIFAR10	CUB200	CARS196	FLICKR25K
TITAN w/o $\mathcal{L}_{SL}$	85.04	17.05	15.82	91.98
TITAN w/o $\mathcal{L}_{C}$	84.33	16.67	15.24	91.23
TITAN w/o M	87.56	18.25	16.13	92.65
TITAN w/o R	87.78	18.05	16.08	92.11
TITAN w/o $\boldsymbol{w}$	88.54	18.77	17.24	93.97
TITAN (full)	**89.01**	**19.11**	**17.64**	**94.33**

the area enclosed by the curve corresponding to our method is significantly larger than the remaining three baselines, which indicates that our method has better retrieval performance. In the curve of topN-precision, our method keeps the highest accuracy rate as the number of retrieved samples increases, especially when the number of returned samples is around 100. Overall, we have the conclusion based on these visualization results that when coming to actual image retrieval, TITAN can achieve promising retrieval performances.

Noise Degree Analysis. We plot the performance trends under two different types of noise variation, especially for the high noise rate in Figure 4. This result indicates that the performance of PRISM decreases rapidly with an increase in noise, as PRISM is unable to construct a clean memory bank at the beginning if the noise ratio is too high. Even with a high noise ratio, our proposed TITAN is still able to deliver the best performance. Due to curriculum learning and blending the semantic information of the prototype with the sample features, we gradually construct a relatively accurate prototype for semantic learning. In general, our TITAN achieves great performance under various noise ratios, demonstrating the robustness and superiority of our proposed TITAN.

Feature Dimension Analysis. Figure 5 shows the MAP@R results of TITAN and PRISM methods implemented on CUB and FLICKR25k with feature dimensions varying from 128 to 2048. It can be found that the MAP@R value has a small decrease in both 128 and 2048 dimensions for TITAN, while for PRISM there is no obvious pattern of change. In general, probably because of the L2 normalization, the MAP@R value does not change much with the increase of hidden dimension. Note that our method outperforms the PRISM for all the cases, which proves the robustness of our method.

4.3. Ablation Study

We present an ablation study to investigate the effectiveness of the important inner module in our proposed TITAN and the experiment results are shown in Table 4. Specifically, we design the following model variants as: (1) **TITAN w/o $\mathcal{L}_{SL}$** removes similarity learning loss for denoising in Eqn. 7. (2) **TITAN w/o $\mathcal{L}_{C}$** removes cluster learning loss for retrieval in Eqn. 11. (3) **TITAN w/o M** removes the sampled virtual feature in Eqn. 4. (4) **TITAN w/o R** removes the retrieval-based refinement procedure. (5) **TITAN w/o $\boldsymbol{w}$** removes confidence scores in Eqn. 8. From these results, we can draw some observations as follows: First, **TITAN w/o $\mathcal{L}_{C}$** has lower MAP@R scores than **TITAN w/o $\mathcal{L}_{SL}$**, perhaps because in the case of a large number of sample categories in the training set, using a small batch size will lead to few positive sample pairs in the mini-batch, which will affect the learning of similarity. Second, **TITAN w/o M** performs worse than TITAN, which indicates that our use of prototypical mixing is effective, it makes the feature representation more robust and alleviates the overfitting of the model to dirty data to some extent. Third, MAP@R values corresponding to **TITAN w/o R** drop by about the same amount as **TITAN w/o M**, which proves that Retrieval-based refinement is an important and influential module in our proposed TITAN, it provides an effective way to refine noisy samples for pairwise similarity learning framework in retrieval tasks. Finally, the slightly lower performance of **TITAN w/o $\boldsymbol{w}$** compared to TITAN suggests that confidence scores can improve performance by attenuating the impact of unreliable mixed samples.

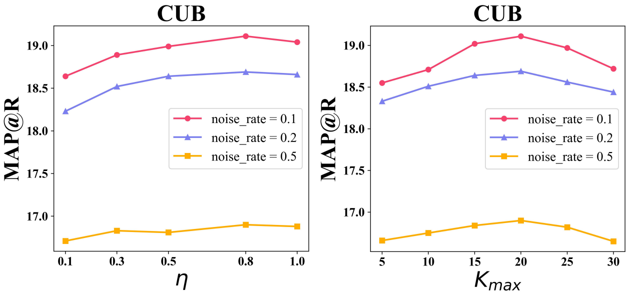

Figure 6. Sensitivity analysis of two hyper-parameters on CUB.

4.4. Sensitivity Analysis

We investigate the effect of hyper-parameters η and K_{max} on model retrieval performance on CUB with different noise rates. The loss balance coefficient η controls the weight between different losses, firstly we vary η in [0.1,0.3,0.5,0.8,1.0] with other parameters fixed, Figure 6 illustrates the experimental results, we can find that performance of our method is not sensitive to η when it varies from 0.5 to 1.0, which demonstrates that the convergence of our proposed algorithm is stable. Moreover, we fix all other parameters and change K_{max} in [5,10,15,20,25,30], which controls the max retrieval number of each prototype, as shown in Figure 6, the MAP@R results decrease when K_{max} is set in [25,30], the potential reason is that a high K_{max} will cause more samples to change their labels, but our constructed prototype is not completely reliable at the early stage of training, this leads to refinement which may in turn increase the noise rate. From the analysis, we set η

as 0.8 and K_{max} as 20 respectively for the proposed TITAN as default.

4.5. Case Study

We present the top ten images retrieved by PRISM and the proposed TITAN on FLICKR25K. As depicted in Figure 7, the examples enclosed in green boxes represent accurate results, whereas those enclosed in red boxes are wrong results. Additionally, we have labeled the category of each image below it. It is evident that the proposed approach yields significantly more relevant and precise retrieval outcomes for the given queries, demonstrating the superiority of our methodology in handling realistic scenarios.

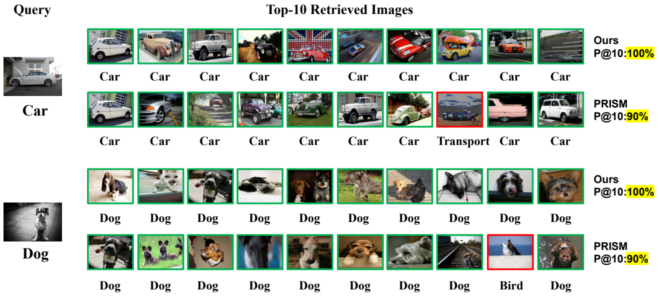
Query
Top-10 Retrieved Images
Car
Car Car Car Car Car Car Car Car Car Car
Ours P@10:100%
Car Car Car Car Car Car Car Transport Car Car
PRISM P@10:90%
Dog
Dog Dog Dog Dog Dog Dog Dog Dog Dog Dog
Ours P@10:100%
Dog Dog Dog Dog Dog Dog Dog Dog Bird Dog
PRISM P@10:90%

Figure 7. Example of the Top10 returned images with 512-dimensional feature on FLICKR25K.

4.6. Visulaization Analysis

T-SNE Visualization. We adopt T-Distributed Stochastic Neighbor Embedding (T-SNE) [39] to compress the dense feature vectors into the two-dimensional plane to visualize the correlations between the features embedded and their one-hot labels, which come from the feature space and the label space respectively. Figure 8 illustrates the visualization of dense vectors generated by our method and the competitive baselines, i.e., PRISM and Jo-SRC. Compared with the two baselines, our method can generate actually compact clusters and can separate each category more clearly while at the same time bringing feature vectors with similar semantic information closer together. This visualization shows that our method can generate dense features that are more discriminative, which could facilitate better image retrieval.

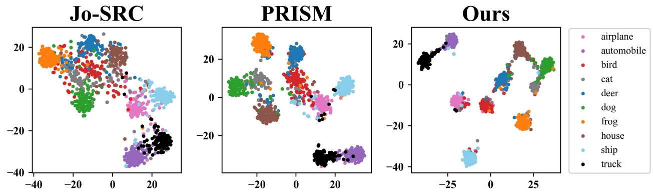

Figure 8. The t-SNE visualization of 512-dimensional feature vectors on CIFAR10.

5. Conclusion

This paper studies a practical but less-explored problem of label noise-resistant image retrieval and proposes a novel method TITAN, which simultaneously corrects label noise and mitigates the impacts of memorization to solve the problem. We describe different prototypes with Gaussian distributions in the hidden space, which directs Mixing to generate synthetic samples. In addition, we collect similar samples for each prototype ranging from easy to hard during optimization, which refines noisy samples in a precise and class-balanced way. Extensive experiments on five benchmark datasets reveal that our proposed TITAN can outperform various state-of-the-art methods. In future works, we would extend our TITAN to more scenarios such as learning to hash and cross-modal retrieval.

References

[1] Artem Babenko and Victor Lempitsky. Aggregating local deep features for image retrieval. In *ICCV*, pages 1269–1277, 2015. 1, 2

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, pages 41–48, 2009. 2, 5

[3] 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. 5

[4] 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. 5

[5] Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. In *NeurIPS*, 2000. 2, 4

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*, 2020. 4

[7] Wei Chen, Yu Liu, Weiping Wang, Erwin M Bakker, Theodoros Georgiou, Paul Fieguth, Li Liu, and Michael S Lew. Deep learning for instance retrieval: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[8] Ching-Yao Chuang, R Devon Hjelm, Xin Wang, Vibhav Vineet, Neel Joshi, Antonio Torralba, Stefanie Jegelka, and Yale Song. Robust contrastive learning against noisy views. In *CVPR*, pages 16670–16681, 2022. 2

[9] Enyan Dai, Charu Aggarwal, and Suhang Wang. Nrgnn: Learning a label noise resistant graph neural network on sparsely and noisily labeled graphs. In *KDD*, 2021. 2

[10] Lei Feng, Senlin Shu, Zhuoyi Lin, Fengmao Lv, Li Li, and Bo An. Can cross entropy loss be robust to label noise? In *IJCAI*, 2021. 2

[11] Aritra Ghosh, Himanshu Kumar, and P Shanti Sastry. Robust loss functions under label noise for deep neural networks. In *AAAI*, volume 31, 2017. 2

[12] Aritra Ghosh, Naresh Manwani, and PS Sastry. On the robustness of decision tree learning under label noise. In *PAKDD*, pages 685–697, 2017. 1

[13] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *ECCV*, pages 241–257, 2016. 1

[14] Geonmo Gu, Byungsoo Ko, and Han-Gyu Kim. Proxy synthesis: Learning with synthetic classes for deep metric learning. In *AAAI*, volume 35, pages 1460–1468, 2021. 1, 2, 4

[15] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018. 2, 3, 4

[16] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Discriminative deep metric learning for face verification in the wild. In *CVPR*, pages 1875–1882, 2014. 1, 2

[17] Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, 2008. 5

[18] 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. 2, 3, 6

[19] Jayashree Kalpathy-Cramer, Alba García Seco de Herrera, Dina Demner-Fushman, Sameer Antani, Steven Bedrick, and Henning Müller. Evaluating performance of biomedical image retrieval systems—an overview of the medical image retrieval task at imageclef 2004–2013. *Computerized Medical Imaging and Graphics*, 39:55–61, 2015. 1

[20] Nazmul Karim, Mamshad Nayeem Rizve, Nazanin Rahnavard, Ajmal Mian, and Mubarak Shah. Unicon: Combating label noise through uniform selection and contrastive learning. In *CVPR*, pages 9676–9686, 2022. 2

[21] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *CVPR*, pages 3238–3247, 2020. 1, 2, 4, 5

[22] 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. 5

[23] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[24] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *ICLR*, 2021. 2

[25] Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *AISTAT*, 2020. 1

[26] Chang Liu, Kunpeng Li, Michael Stopa, Jun Amano, and Yun Fu. Discovering informative and robust positives for video domain adaptation. In *ICLR*, 2022. 4

[27] Chundi Liu, Guangwei Yu, Maksims Volkovs, Cheng Chang, Himanshu Rai, Junwei Ma, and Satya Krishna Gorti. Guided similarity separation for image retrieval. In *NeurIPS*, 2019. 2

[28] 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. 2, 3, 5, 6

[29] Aditya Krishna Menon, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Can gradient clipping mitigate label noise? In *ICLR*, 2020. 1

[30] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *ECCV*, pages 681–699, 2020. 1, 2

[31] Duc Tam Nguyen, Chaithanya Kumar Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Laura Beggel, and Thomas Brox. Self: Learning to filter noisy labels with self-ensembling. In *ICLR*, 2020. 2

[32] Youngmin Oh, Beomjun Kim, and Bumsub Ham. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation. In *CVPR*, 2021. 2

[33] Shichao Pei, Lu Yu, Guoxian Yu, and Xiangliang Zhang. Graph alignment with noisy supervision. In *WWW*, pages 1104–1114, 2022. 5

[34] Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. Identifying mislabeled data using the area under the margin ranking. In *NeurIPS*, pages 17044–17056, 2020. 5

[35] Qi Qian, Lei Shang, Baigui Sun, Juhua Hu, Hao Li, and Rong Jin. Softtriple loss: Deep metric learning without triplet sampling. In *ICCV*, pages 6450–6458, 2019. 1, 2

[36] Karsten Roth, Oriol Vinyals, and Zeynep Akata. Non-isotropy regularization for proxy-based deep metric learning. In *CVPR*, pages 7420–7430, 2022. 1, 2, 4

[37] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Robust learning by self-transition for handling noisy labels. In *KDD*, pages 1490–1500, 2021. 2

[38] 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. 4, 6

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

[40] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *ECCV*, pages 268–285, 2020. 5

[41] Vladimir Vapnik and Vlamimir Vapnik. Statistical learning theory wiley. *New York*, 1(624):2, 1998. 4

[42] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 5

[43] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *CVPR*, pages 1386–1393, 2014. 1, 2

[44] Jingdong Wang, Ting Zhang, Nicu Sebe, Heng Tao Shen, et al. A survey on learning to hash. 40(4):769–790, 2017. 1, 2

[45] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *CVPR*, pages 5022–5030, 2019. 1, 2

[46] Xinshao Wang, Yang Hua, Elyor Kodirov, Guosheng Hu, Romain Garnier, and Neil M Robertson. Ranked list loss for deep metric learning. In *CVPR*, pages 5207–5216, 2019. 1, 2

[47] Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia. Iterative learning with open-set noisy labels. In *CVPR*, pages 8688–8696, 2018. 1

[48] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *CVPR*, 2020. 2

[49] Tobias Weyand and Bastian Leibe. Discovering favorite views of popular places with iconoid shift. In *ICCV*, pages 1132–1139, 2011. 1

[50] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *ICCV*, pages 2840–2848, 2017. 1, 2

[51] Songhua Wu, Xiaobo Xia, Tongliang Liu, Bo Han, Mingming Gong, Nannan Wang, Haifeng Liu, and Gang Niu. Class2simi: A noise reduction perspective on learning with noisy labels. In *ICML*, 2021. 4

[52] 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. 5

[53] Zhibo Yang, Muhammet Bastan, Xinliang Zhu, Douglas Gray, and Dimitris Samaras. Hierarchical proxy-based loss for deep metric learning. In *WACV*, pages 1859–1868, 2022. 1, 2, 4

[54] 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, 5

[55] Luca Zancato, Alessandro Achille, Tian Yu Liu, Matthew Trager, Pramuditha Perera, and Stefano Soatto. Train/test-time adaptation with retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15911–15921, 2023. 2

[56] Yu-Wei Zhan, Yongxin Wang, Yu Sun, Xiao-Ming Wu, Xin Luo, and Xin-Shun Xu. Discrete online cross-modal hashing. *Pattern Recognition*, 122:108262, 2022. 1, 2

[57] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 2021. 2, 3

[58] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 2

[59] Jing Zhang, Jianwen Xie, and Nick Barnes. Learning noise-aware encoder-decoder from noisy labels by alternating back-propagation for saliency detection. In *ECCV*, 2020. 2

[60] Yaobin Zhang, Weihong Deng, Yaoyao Zhong, Jiani Hu, Xian Li, Dongyue Zhao, and Dongchao Wen. Adaptive label noise cleaning with meta-supervision for deep face recognition. In *ICCV*, pages 15065–15075, 2021. 2

[61] Guoqing Zheng, Ahmed Hassan Awadallah, and Susan Dumais. Meta label correction for noisy label learning. In *AAAI*, 2021. 2

[62] Liang Zheng, Yi Yang, and Qi Tian. Sift meets cnn: A decade survey of instance retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1224–1244, 2017. 1, 2