Check for updates

Towards Food Image Retrieval via Generalization-oriented Sampling and Loss Function Design

JIAJUN SONG*, Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, China
ZHUO LI*, Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, China
WEIQING MIN, Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, China
SHUQIANG JIANG, Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, China

Food computing has increasingly received widespread attention in the multimedia field. As a basic task of food computing, food image retrieval has wide applications, *i.e.*, food image retrieval can help users to find the desired food from a large number of food images. Besides, the retrieved information can be applied to establish a richer database for the subsequent food content-related recommendation. Food image retrieval aims to achieve better performance on novel categories, and it is thus worth studying to transfer the embedding ability from the training set to the unseen test set, *i.e.*, the generalization of the model. Food is influenced by various factors, such as culture and geography, leading to great differences between domains, such as oriental food and western food. Therefore, it is challenging to study the generalization of the model in food image retrieval. In this paper, we improve the classical metric learning framework and propose a generalization-oriented sampling strategy, which boosts the generalization of the model by maximizing the intra-class distance from a proportion of positive pairs to avoid the excessive distance compression in the embedding space. Considering that the existing optimization process is in an opposite direction to our proposed sampling strategy, we further propose an adaptive gradient assignment policy named gradient-adaptive optimization, which can alleviate the intra-class distance compression during optimization by assigning different gradients to different samples. Extensive evaluation on three popular food image datasets demonstrates the effectiveness of the proposed method. We also experiment on three popular general datasets to prove that solving the problem from the generalization can also improve the performance of general image retrieval. Code is available at https://github.com/Jiajun-ISIA/Generalization-oriented-Sampling-and-Loss

CCS Concepts: • **Computing methodologies** → **Visual content-based indexing and retrieval**.

Additional Key Words and Phrases: Food computing, Image retrieval, Deep learning

*Equal contribution.

Authors' addresses: Jiajun Song, jiajun.song@vipl.ict.ac.cn, Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, No.6 Kexueyuan South Road, Beijing, China; Zhuo Li, zhuo.li@vipl.ict.ac.cn, Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, No.6 Kexueyuan South Road, Beijing, China; Weiqing Min, Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, No.6 Kexueyuan South Road, Beijing, China, weiqingmin@ict.ac.cn; Shuqiang Jiang, sqjiang@ict.ac.cn, Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, No.6 Kexueyuan South Road, Beijing, China.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
© 2023 Copyright held by the owner/author(s).
1551-6857/2023/5-ART
https://doi.org/10.1145/3600095

Just Accepted

1 INTRODUCTION

Food is closely related to the healthy life of human beings and food related research thus becomes a hotspot. Food image retrieval aims to find all relevant images from the original food database through a query, and it is an important issue in the multimedia field. As an important branch of content-based image retrieval (CBIR), the development of food image retrieval expands the scale of CBIR and announces that this field has made great progress. Therefore, food image retrieval has high academic value. Besides, food image retrieval has higher application value. First, food image retrieval can help consumers with special needs to find the desired food from a large number of food images. For example, fitness coaches and body managers usually need to use food-related search engines to find healthy meals from a wide range of dishes displayed by a large number of stores on the Internet. Food image retrieval can help them to find the best match from a large number of food images. Second, food image retrieval can solve the problem of dynamic categories which can not be solved by common food recognition paradigm. There are many kinds of unlabeled novel foods and labelling these food images costs a lot, while food image recognition requiring a set of annotated images to train a robust classifier. Food image retrieval can be used to find similar foods among the available ones and suggest the most possible food class to solve the problem. Third, food image retrieval can be used to enrich the database of food recommendation system. The catering-oriented social media, such as Yelp[1] and Meituan[2], utilizes the dish name-based recommendation system to predict the users' preference. However, it usually fails because different restaurants may give different names to the same dish and the recommendation system can not match the corresponding dish name in the database. The food image-based recommendation system can avoid the problem. Therefore, food image retrieval helps to recommend food more accurately by enriching the database, which avoids the matching failure caused by the same dish with different names.

With the rapid development of the food field, new food categories are produced every day. Existing methods [10, 15, 44] are not suitable for retrieving corresponding images for these unseen categories. Therefore, we focus on food image retrieval on unseen categories, which is also consistent with general image retrieval in current multimedia field. The generalization of the model to novel food categories is an important issue to solve for food image retrieval. However, there are few works to explore this issue in food image retrieval. Improving the generalization of the model in food image retrieval is to cross the gap among categories to achieve the knowledge transfer, and the challenge is to solve the problem with large inter-class differences of different food domains. Specifically, there are many domains of food, such as oriental food and western food. Western food is often visually represented as a simple combination of processed ingredients, such as steak, fish & chips, and hamburgers. Due to the uniqueness of cooking methods, oriental food such as Chinese food is often processed again by mixing various ingredients, which is inseparable. The situation leads to the great visual difference of food in different domains. As shown in Fig. 1, the Cripsy sweet & sour pork slices on the left differs greatly from the Beefsteak on the right. When two categories belong to the training set and the test set, respectively, the performance may be poor.

To solve this problem, we improve the metric learning framework and propose a generalization-oriented sampling strategy, named ρ Sampling. The previous work [41] indicates that the generalization of the model on representation learning such as image retrieval under considerable shifts between training and testing distribution is hurt by excessive feature compression. Therefore, making features occupy a larger hyperspace is helpful to the improvement of the generalization of models. A simple example is shown in Fig. 2, although two models both successfully separate the data on the training set, the left model fails to separate all test classes due to the excessive compression, leading to the confusion on the test set. On the contrary, the right model successfully separates the test classes by keeping the intra-class variants, which means that the model transfers the discrimination

[1] https://www.yelp.com/
[2] https://meituan.com/

Just Accepted

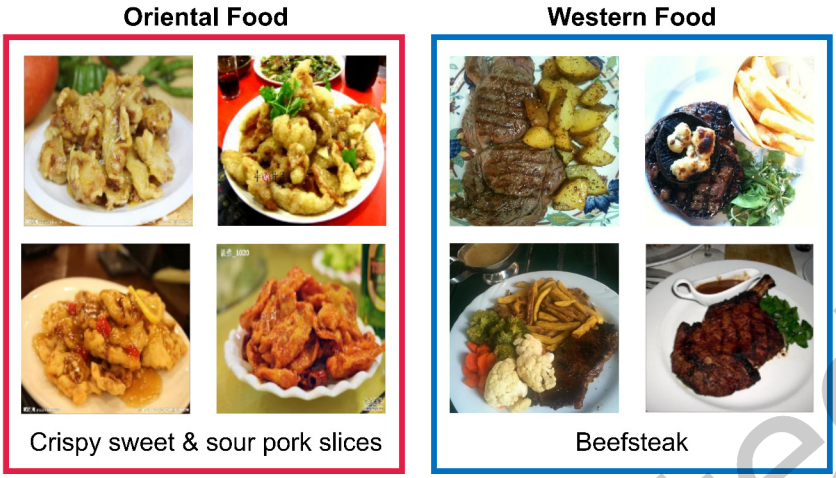

Fig. 1. The difference between two domains of food.

ability from the training set to the test set. Based on this, we argue that the training shouldn't blindly minimize the intra-class distance. It is necessary to maximize the space utilization to achieve better generalization. The proposed sampling strategy avoids excessive compression by maximizing a few intra-class distances during the training, which can take instances from the same class apart to occupy a larger hyperspace to achieve the generalization advancement.

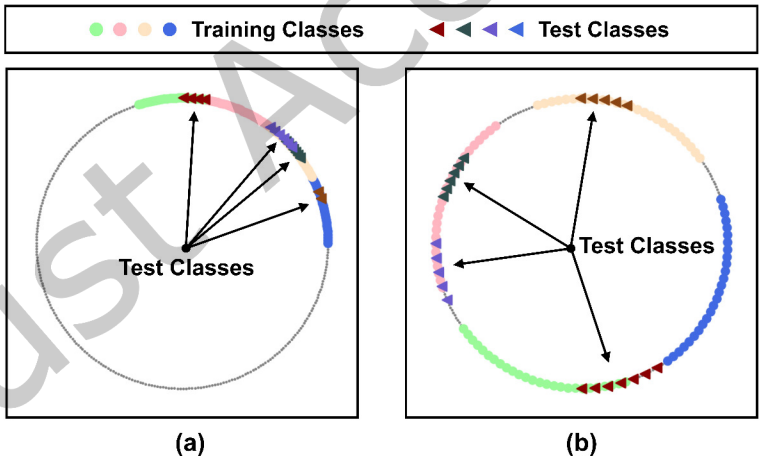

Fig. 2. The toy example for the distribution of the training set and test set. Different colors represent different categories. (a) shows the distribution of the excessive compression. (b) shows the optimized distribution.

Furthermore, after we investigate the widely-used loss functions from the perspective of the optimization under the metric learning framework, we find that recent methods [17, 28] prefer to force all positive instances to be closed during training by minimizing the distance between the anchor and the positive one. This optimization method will force positive instances with different confidence to the anchor position and lead to the excessive

Just Accepted

compression, which is opposite to the above requirement. Therefore, we introduce an adaptive gradient assignment policy, named Gradient-Adaptive Optimization (GAO) to avoid the excessive compression. The proposed GAO keeps the intra-class distance by assigning smaller gradients to the edge sample to avoid the excessive compression and achieves better generalization. The edge sample is the sample with low confidence, which is far from the class center. Meanwhile, GAO can achieve more stable optimization during training because GAO can reduce the influence of edge samples that may be potential noise data. We just verify it on two popular loss functions, contrastive loss and margin loss, and the similar strategy can also be used to other loss functions for performance improvement. Extensive evaluation on three food image datasets demonstrates the effectiveness of the proposed method. In order to prove that solving the problem from the generalization can also improve the performance of other tasks, we conduct the evaluation on three popular general datasets to verify the performance on the general image retrieval task.

In summary, the main contributions of this work are listed below.

- We are the first to investigate food image retrieval on unseen categories and improve the generalization of the model for the performance gain to the best of our knowledge.
- We improve the classical metric learning framework for food image retrieval and propose a generalization-oriented sampling strategy named ρ Sampling and an adaptive gradient assignment policy named Gradient-Adaptive Optimization to avoid the excessive distance compression in the embedding space to improve the generalization of the model.
- We conduct a benchmark on three datasets and experimental results demonstrate that the proposed method achieves the state-of-the-art performance. Experiments on three general datasets prove that the method can be further popularized in general image retrieval.

2 RELATED WORK

Our work is closely related to three research fields: (1) food retrieval, (2) deep metric learning and (3) generalization in image retrieval.

2.1 Food Retrieval

Recently, Min *et al.* [32] provided a comprehensive survey of food retrieval and other food-related works [19, 22, 31, 38, 55, 65]. According to retrieval types, food-relevant retrieval consists of three types: visual food retrieval, recipe retrieval and cross-modal recipe-image retrieval. For food image retrieval, Gianni *et al.* [2] introduced a search engine for restaurant retrieval based on dishes rather than using their categories. Farinella *et al.* [15] conducted the food image retrieval by comparing images through similarity measures, where food images were represented as vectors through the combination of different types of features, such as SIFT and Bag of Textons. For recipe retrieval, Wang *et al.* [54] investigated the underlying features of Chinese recipes. Based on workflow-like cooking procedures, they modelled recipes as graphs and further proposed a novel similarity measurement based on frequent patterns, and devised an effective filtering algorithm to support efficient online searching. Recently, Chang *et al.* [6] proposed an interactive system RecipeScape to analyze multiple recipes for one dish. They changed the recipe instruction into a tree-structure representation for recipe similarity calculation. Xie *et al.* [59] further jointly utilized various features such as cooking flow features, eating features and nutrition features to create a hybrid semantic item model for recipe search. Besides food/recipe retrieval, there are some researches[9, 29, 33, 53] on cross-modal recipe-image retrieval. Salvador *et al.* [42] firstly proposed a large-scale, structured corpus of over one million cooking recipes and food images, namely Recipe1M. Micael *et al.* [29] extended previous works by providing a double-triplet strategy to jointly express both the retrieval loss and the classification one for cross-modal retrieval. Chen *et al.* [9] further used attention mechanism to model the recipe and align the word embedding of a recipe with its image feature.

Just Accepted

Compared to previous works, the goal of our task is to search the most similar image from a number of images, which is different from other food retrieval tasks, i.e., recipe retrieval and cross-modal recipe-image retrieval. Besides, different from existing methods which are suitable for seen categories, our method focuses on unseen categories and improves the generalization of the model on these categories.

2.2 Deep Metric Learning

Deep metric learning methods [30, 56, 58, 63] aim to learn a non-linear transformation to project input images into an embedding space, in which images with similar semantic information have larger similarity and vice versa. Pairwise loss functions represented by contrastive loss [21, 43] were widely used in operating on image pairs. These loss functions maximized the similarity between positive instances while encouraging a margin between negative pairs. Different from above works, triplet networks have been proposed to solve the ranking problem [1, 37, 61] via optimizing the relative distance between positive and negative pairs. However, Cui *et al.* [11] argued that there are too many pairs or tuples which have no effect in updating a model due to their poor discriminative power. They realized that these triplets wasted much time and resources, therefore slowing down the training and leading to worse performance. To this end, instead of selecting random samples, a series of sampling strategies were proposed to mine hard examples [11, 36, 43]. Online mining and offline mining are two strategies to mine hard examples in the dataset. Online hard example mining usually mines hard pairs (triplets) in a training batch during training, while offline mining focusing on the whole dataset or a subset of the whole dataset after each epoch. Online mining is usually faster than offline hard example mining, while offline hard example mining can consider more samples and capture more realistic data distribution.

In this paper, we follow the classical deep metric learning framework and further propose a novel sampling strategy and a loss function design to improve the generalization of the model. Specifically, we propose the ρ Sampling and GAO to boost the generalization by avoiding the excessive compression within the class.

2.3 Generalization in Image Retrieval

Recent works [7, 13, 25, 48, 60, 64] widely used the metric learning method for the image retrieval task, and a series of sampling methods and loss functions have further proposed to improve the performance of image retrieval. Tishby *et al.* [49] studied the concept of compression to explain such performance gain on basis of the learned embedding space for the metric learning method. The recent work [51] introduced Singular Value Decomposition (SVD) on the data representations and obtained an increased decay of singular values, which linked the compression to the class-conditioned flattening of representation. Thus, class representations occupied a more compact volume, thereby reducing the number of directions with significant variance. For classic classification scenarios with i.i.d. training and test distributions, *i.e.*, the overlap setup, researchers strong focused on the most discriminative directions. However, this overly discarded features which could capture data characteristics outside the training distribution, which damaged the generalization of the model. Bellet *et al.* [3] indicated that generalization in the no-overlap problem like metric learning was hindered due to the shift in the training and test distribution. Based on this, Roth *et al.* [41] proposed to retain a considerable amount of directions of significant variance (DoV) to learn a well generalizing embedding function. Therefore, the metric learning under considerable shifts between training and testing distribution was hurt by the excessive feature compression.

Inspired by the work [41], we aim to further avoid excessive compression to improve the generalization of the model. The proposed ρ sampling boosts the generalization of the model by maximizing the intra-class distance from a proportion of positive pairs to avoid excessive distance compression in the embedding space and GAO further alleviates the intra-class distance compression during optimization by assigning different gradients to different samples.

Just Accepted

3 OUR METHOD

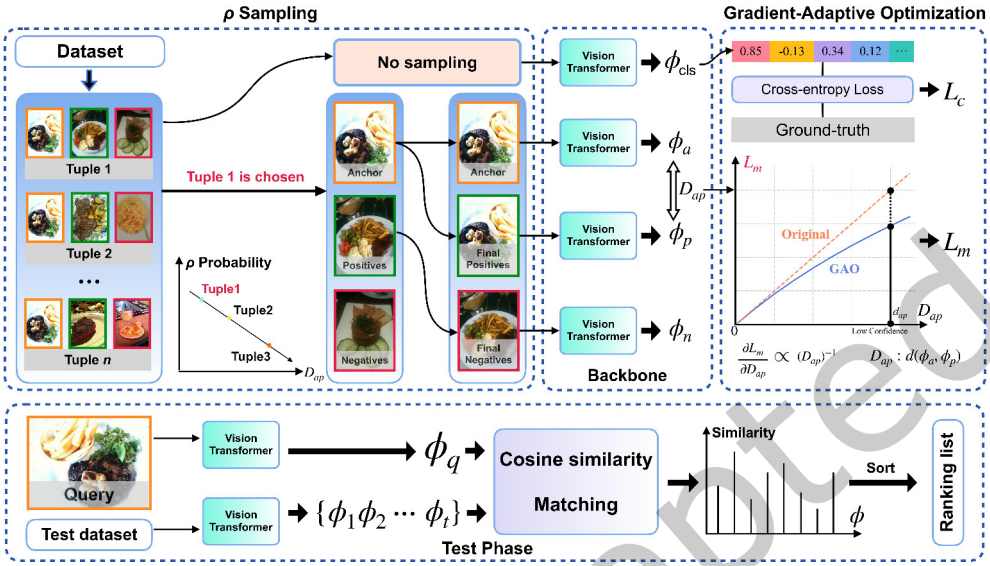

Fig. 3. The proposed food image retrieval framework. ϕ_a, ϕ_p and ϕ_n are image descriptors and ϕ_{cls} is the logit for classification.

As shown in Fig. 3, we improve the classical metric learning framework by introducing the novel sampling strategy and loss function to avoid the excessive distance compression in the embedding space and improve the generalization. Particularly, we propose the ρ Sampling and GAO, which maintain the intra-class distance and achieve better generalization performance, which will be detailed in Sec. 3.1 and 3.2, respectively. Without loss of generality, we choose vision transformer (ViT) [14] as the backbone. In the test phase, all images in the test set are fed into the backbone to obtain image descriptors to compute cosine similarities among the query and other images. Finally, we sort these similarities and obtain the final ranking list.

3.1 ρ Sampling for Generalization

Food image retrieval aims to achieve better performance on unseen classes based on the model trained via seen classes. Therefore, generalization becomes crucial. As mentioned above, the generalization of the model on representation learning such as image retrieval under considerable shifts between training and testing distribution is hurt by the excessive distance compression in the embedding space. Due to the obvious differences between different domains of food images, it is more challenging to study the generalization in food image retrieval. In this paper, we argue that in order to achieve better generalization performance, the training should not only minimize the intra-class variance and inter similarity but also distribute the data into wider space. As mentioned above, the model which makes full use of the hyperspace is considered to have better generalization. Such model achieves the ability to distribute unseen classes onto hyperspace uniformly and widely, and avoids the confusion of the test data.

To this end, we propose a novel sampling strategy namely ρ Sampling, which keeps the intra-class variance by randomly maximizing some intra-class distances during training. Specifically, in order to avoid the excessive compression, an intuitive idea is to relax the optimization of positive samples. That is because metric learning methods make the distance between the anchor and the positive instance closer, which leads to the excessive

Just Accepted

distance compression in the embedding space. To solve the problem, we randomly select a part from all tuples $T = (x_{a_1}, x_{p_1}, x_{n_1}), (x_{a_2}, x_{p_2}, x_{n_2}), ..., (x_{a_k}, x_{p_k}, x_{n_k})$ for relaxation and we assign the same ρ probability for each tuple to approximately obtain ρ ratio of all tuples for relaxation, where x_a is the anchor, x_p is the positive instance, x_n is the negative instance and k is the size of the batch. For these selected tuples, we replace positive instances x_p with anchors x_a, and optimize original positive instances as negative ones, which is defined as:

$$t = (x_a, x_p, x_n) \rightarrow t' = (x_a, x_a, x_p) \tag{1}$$

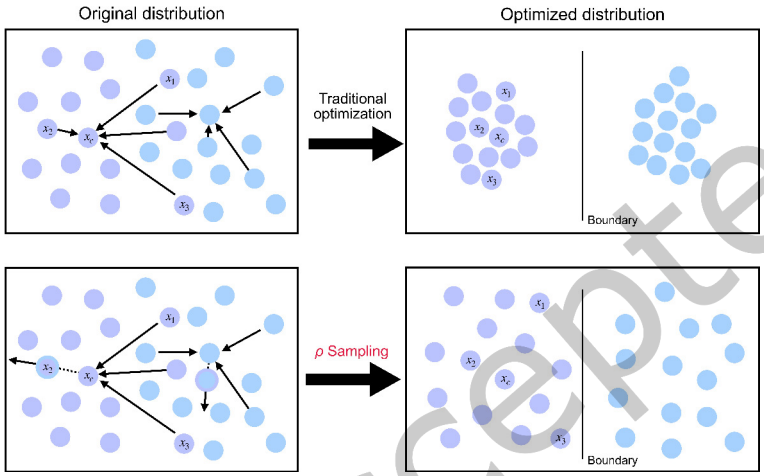

Fig. 4. Visualization of the optimization process. Circles of different colors represent instances of different food classes. According to the ρ Sampling, the positive x_2 is replaced as negative ones to avoid over-compression for better generalization. The dashed arrows represent the optimization processes in training.

This strategy pushes samples from the same class apart, enabling the model to capture extra discriminative features and result in outstanding generalization. As shown in Fig. 4, the proposed ρ Sampling avoids the excessive distance compression in the embedding space and makes full use of the hyperspace to have better generalization. However, it may be unreasonable to randomly select positive instances to change. After ensuring that the proportion of selected tuples is ρ, we hope to select more appropriate tuples for this inverse optimization. Therefore, we analyse the distribution of pairwise distance, which follows

$$q(d) \propto d^{n-2}\left[1 - \frac{1}{4}d^2\right]^{\frac{n-3}{2}} \tag{2}$$

where d is the distance between two embeddings and n is the dimension of the embedding. In high dimensional space, $q(d)$ approaches $\mathcal{N}\left(\sqrt{2}, \frac{1}{2n}\right)$, which means that if we randomly select training samples, the distances between the positive pairs will usually be $\sqrt{2}$, *i.e.*, hard positive pairs. As shown in Fig. 4, if hard positive samples (x_1 and x_3) are used for inverse optimization, these edge samples will be further far from the center, which leads to the confusion between classes. Although this method maintains the intra-class variance, the potential confusion between categories weakens the discrimination between classes. We believe that occupying a larger hyperspace should ensure the boundary. Based on this, we should give priority to samples which are close to the center to keep intra-class variance. Because that such samples are confident enough and the influence of their change is

Just Accepted

less. To this end, we define the probability of a tuple being selected as,

$$P(t^*|a,p) \propto (D_{ap})^{-1} \tag{3}$$

where D_{ap} is the distance between the anchor and the positive instance $d(\phi_a, \phi_p)$. In other words, we prefer to select easy positive pairs as altered samples.

Discussion. The conventional metric learning method makes the distance between samples of the same category closer and pushes samples of different categories away. Although this method can distinguish samples of different categories, it will inevitably lead to the excessive distance compression of the same kind of samples in the embedding space. Because this method narrows the distance between all positive samples equally without considering the distribution of a category. As shown in Fig. 4, the traditional optimization leads to the excessive distance compression in the embedding space. On the contrary, the proposed ρ Sampling maximizes the intra-class distance from a proportion of positive pairs to avoid the excessive compression to improve the generalization of the model.

3.2 Gradient-Adaptive Optimization (GAO)

In order to further reduce the excessive distance compression in the embedding space in the class and improve the generalization performance of the model, we propose an adaptive gradient assignment policy named Gradient-Adaptive Optimization (GAO) to cooperate with the proposed ρ Sampling. It is beneficial to the generalization performance of the model by paying less attention to hard positive instances and slowly correcting them, which occupies a larger hyperspace. Based on this, the strategy slowly corrects instances with lower confidence by giving them lower gradients. To this end, we define the strength of a positive instance during the optimization as:

$$s(x_p) \propto (D_{ap})^{-1} \tag{4}$$

The equation means that the optimization strength of the positive instance decreases with the increase of D_{ap}.

Since derivative functions of distances between anchors and positive instances in original margin loss and contrastive loss are constant functions, which are convenient for us to study the impact of different gradient allocations on the retrieval performance. Therefore, we apply the proposed GAO to these two loss functions to further verify the effectiveness of the strategy.

Without loss of generality, we take contrastive loss as an example. It is formulated as:

$$L_{m_1} = D_{ap} + \max(m - D_{an}, 0) \tag{5}$$

where m is the margin. Contrastive loss forces all positive instances to be close during training by minimizing the distance between the anchor and the positive one. Meanwhile, contrastive loss will assign equal strength to different instances and synchronously correct them, shown in Fig. 5 (b). The gradient of D_{ap} keeps constant when D_{ap} changes, which means that positive instances with different confidence will be under the same optimization strength.

The contrastive loss combined with GAO is defined as follows:

$$L'_{m_1} = \log(1 + D_{ap}) + \max(m - D_{an}, 0) \tag{6}$$

L'_{m_1} and its derivative function with respect to D_{ap} is shown in Fig. 5 (b). When D_{ap} is low, the gradient of D_{ap} is large, which means strong optimization strength. In contrast, when D_{ap} is large, the gradient of D_{ap} is small, which means weak optimization strength. The improved contrastive loss will quickly correct easy samples (with high confidence) while slowly correcting edge samples. Such strategy avoids the excessive compression and weakens the impact of wrong correction caused by the potential noise data and achieves more stable optimization.

Just Accepted

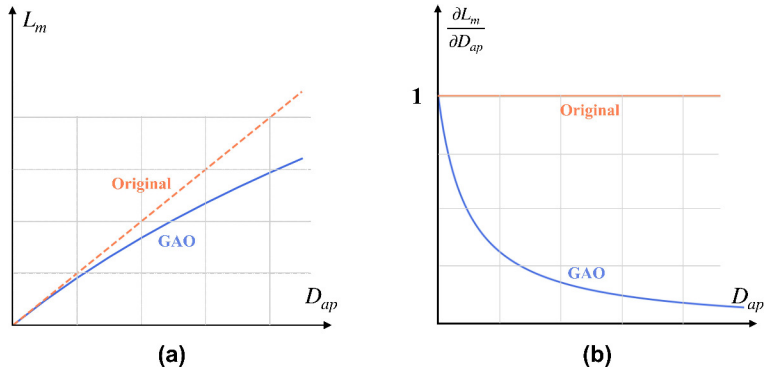

Fig. 5. Compared to the original contrastive loss, the proposed training strategy assigns smaller gradient to harder positives and slowly correct them to achieve the adaptive optimization.

The proposed strategy has also been used to improve margin loss because the derivative function of distances between anchors and positive instances is the constant function. The original margin loss is defined as,

$$L_{m_2} = \gamma + \{D_{ap} - \beta\} - \{D_{an} - \beta\} \tag{7}$$

where γ is the triplet margin. The original margin loss extends the standard triplet loss by introducing a dynamic, learnable boundary β between positive and negative pairs, which transfers the common triplet ranking problem to a relative ordering of pairs. The improved margin loss is defined as,

$$L'_{m_2} = \gamma + \{\log(1 + D_{ap}) - \beta\} - \{D_{an} - \beta\} \tag{8}$$

Similar to the improved contrastive loss, the improved margin loss redistributes the gradient of D_{ap} to quickly correct easy examples while slowly correcting edge samples, which avoids the excessive compression to improve the generalization of the model.

Finally, we add a classification sub-network, leading to the joint optimization on both the metric learning and classification loss. We adopt the classification loss to improve the discrimination on different classes. Features further occupy a larger hyperspace to improve the generalization of the model. The final loss is defined as

$$L = \alpha L_m + (1 - \alpha) L_c \tag{9}$$

where α is an adjustable hyper-parameter, L_m and L_c represent the metric learning loss and classification loss, respectively.

4 EXPERIMENT

Sec. 4.1 shows datasets for evaluation and their division. Sec. 4.2 and Sec. 4.3 shows details of adopted evaluation metrics and the experimental setup, respectively. In Sec. 4.4, we introduce results of the comparison with state-of-the-art, different loss functions and different sampling strategies. We also conduct the cross-domain evaluation experiment to further verify the generalization ability of our method and the impact of hyper-parameters is also shown. Sec. 4.5 shows the qualitative analysis. Finally, we introduce results of the comparison with the state-of-the-art on general image retrieval in Sec. 4.6 to further verify the effectiveness of our method.

4.1 Dataset

We use the following three popular food datasets for evaluation.

Just Accepted

ETH Food-101 [4] is a western food dataset, which contains 101,000 images from 101 categories. Following the test protocol [44], we use the first 70 classes as the train set and the rest 31 classes to evaluate models.

Vireo Food-172 [8] contains 110,241 images from 172 categories, which mainly consists of oriental dishes. The first 120 classes are used for training and the rest 52 classes for testing.

ISIA Food-500 [34] is a large miscellaneous dataset with 399,726 images from 500 classes, which consists of oriental food and western food. We keep the first 350 classes for training, and the last 150 classes for testing.

4.2 Evaluation Metrics

As shown in Fig. 3, we show the pipeline of the test phase. For the test set of all benchmark datasets, every instance is used as the query I_q and the remaining instances are used as the gallery. We use Recall@K (R@K) and MAP@K as main evaluation metrics and Normalized Mutual Information (NMI) to evaluate the generalization of the model. MAP@K includes MAP@1K and MAP@C, and K is the largest number of samples in each class. NMI is defined as the mutual information between cluster labels and the ground truth. The dists@intra (intra-class distance) and dists@inter (inter-class distance) [41] are further used to measure the space occupied by datasets.

4.3 Experimental Setup

We apply the platform proposed by the work [41] to ensure the fairness of experiments. During training, we use vision transformer (ViT) models pretrained on ImageNet [14] and fine-tune them on benchmark datasets. Models are optimized by Adam [23]. The initial learning rate is set to 10^{-6} and the weight decay is 4×10^{-4}. Batch-size is set to 112. To form the mini-batch, we randomly sample k classes and $|P|$ samples per class, where $|P|$ is set to 2. We resize images to 256×256, and randomly crop them to 224×224 as input. During training, we use random crops and random flipping ($p = 0.5$) for data augmentation, and a single center crop of 224×224 is used during evaluation. The embedding size of all models is 128-D. We use the same random seed 0 to avoid the seed-based performance fluctuations.

4.4 Result Analysis

4.4.1 Comparison with State-of-the-art. Table 1 and Table 2 show the performance of recent methods and ours on three benchmark datasets. We observe that our method achieves the state-of-the-art results. In particular, our model outperforms other methods on all evaluation metrics among all datasets. Specifically, our fusion method outperforms others by 0.3-2.6% on R@K, 0.5-1.1% on MAP@K and 1.1-2.6% on NMI. Compared to baseline methods, *i.e.*, contrastive loss and margin loss functions, the proposed method achieves almost 0.3-5% on Recall@1, 0.5-4.8% on MAP@1K and 1-7% on NMI. It shows that the proposed method not only achieves better retrieval performance, but also better generalization performance. It is noteworthy that MAP@1K and MAP@C are the same on ETH Food-101, because each category of them contains 1,000 samples.

4.4.2 Performance of Different Loss Functions. We compare the performance of the pre-trained model and 12 different loss functions under the same setting. Table 1 also summarizes the results of different loss functions into three types. Specifically, as for ranking-based methods, margin loss achieves best performance on ETH Food-101, Vireo Food-172 and ISIA Food-500, which outperforms others about 0.5-1% on Recall@1 and 0.2-1.2% on MAP@1K. We further compare three proxy-based methods and find that the Softmax method which is used for classification tasks achieves competitive performance. We also compare the recent proposed AP-based method, Smooth-AP [5] for comparison. In summary, we find that there is no obvious performance divergence between different types of loss functions on food image retrieval tasks. Interestingly, we find that the recent methods, such as the Circle loss [47] behaves well on general image retrieval tasks, but the performance on food datasets is poor. The probable reason is that such local optimization method leads to poor generalization of the model, which further verifies the importance of generalization in food image retrieval.

Just Accepted

Table 1. Comparison of R@K and NMI performance for all loss functions (%).

Methods		ETH Food-101				Vireo Food-172				ISIA Food-500			
		R@1	R@2	R@4	NMI	R@1	R@2	R@4	NMI	R@1	R@2	R@4	NMI
ImageNet[12](Pre-train)		60.35	72.02	81.52	44.10	68.71	77.00	83.91	44.82	49.86	60.05	69.50	40.13
Pair-based	Contrastive[44]	71.43	80.96	87.79	54.20	82.20	87.97	92.07	63.52	63.25	72.81	80.58	53.57
	Triplet[1]	71.75	81.19	87.83	54.89	82.50	88.21	92.30	64.34	62.97	72.55	80.34	53.52
	N-Pair[45]	65.48	75.93	83.82	46.34	73.02	80.37	86.04	49.53	47.86	58.35	67.97	40.24
	Hist.[50]	72.78	82.01	88.38	52.83	80.78	86.92	91.29	62.04	61.00	71.06	79.26	52.51
	Margin[28]	73.06	82.13	88.75	56.44	82.85	88.30	92.23	65.35	64.62	73.90	81.37	54.97
	Lifted[18]	64.46	75.33	83.48	45.27	71.92	79.26	85.28	47.76	45.24	55.72	65.64	38.32
	Circle[47]	61.11	73.56	83.24	47.11	73.53	81.49	88.10	58.72	49.13	59.21	68.70	39.81
Proxy-based	Softmax[62]	73.14	82.58	88.87	57.45	82.01	87.77	91.95	64.88	67.16	76.40	83.53	59.17
	ProxyNCA[35]	69.99	79.65	86.61	53.49	79.85	85.89	90.63	60.56	66.36	75.76	82.96	57.93
	SoftTriple[39]	63.89	74.76	82.78	44.80	73.99	81.28	87.00	51.29	51.64	61.51	70.46	40.76
	Proxy Synthesis[16]	70.72	80.69	87.80	55.05	81.17	86.66	90.76	63.32	66.98	76.32	83.42	58.90
AP-based	Smooth-AP[5]	71.55	80.74	87.51	53.95	81.95	87.84	91.75	64.11	64.22	73.66	81.08	55.60
	PNP[26]	71.51	80.61	87.30	53.52	82.05	87.54	91.61	64.13	63.61	73.07	80.60	54.63
Ours (Margin+ρ Sampling)		73.52	82.56	88.98	57.57	82.94	88.38	92.31	65.88	64.67	74.13	81.65	57.25
Ours (Margin+ρ Sampling+GAO)		**74.10**	**83.20**	**89.28**	**58.58**	**83.13**	**88.49**	**92.32**	**66.21**	**69.70**	**78.77**	**85.35**	**61.71**

Table 2. Comparison of MAP@K performance for all loss functions (%).

Methods		ETH Food-101		Vireo Food-172		ISIA Food-500	
		MAP@1K	MAP@C	MAP@1K	MAP@C	MAP@1K	MAP@C
ImageNet(Pre-train)		15.12	15.12	13.14	12.08	5.24	5.27
Pair-based	Contrastive[44]	22.53	22.53	31.04	28.49	12.79	12.82
	Triplet[1]	22.92	22.92	31.45	28.87	12.56	12.58
	N-Pair[45]	19.31	19.31	19.41	17.83	5.63	5.66
	Hist.[50]	23.17	23.17	30.59	27.96	12.21	12.23
	Margin[28]	23.38	23.38	31.79	29.17	12.79	12.82
	Lifted[18]	18.41	18.41	18.09	16.62	4.78	4.80
	Circle[47]	15.04	15.04	22.35	20.86	4.11	4.09
Proxy-based	Softmax[62]	23.88	23.88	30.89	28.01	15.35	15.37
	ProxyNCA[35]	21.65	21.65	30.13	27.57	15.89	15.92
	SoftTriple[39]	18.11	18.11	18.65	17.03	5.97	5.98
	Proxy Synthesis[16]	22.03	22.03	26.99	24.52	16.52	16.56
AP-based	Smooth-AP[5]	23.24	23.24	31.78	29.11	14.67	14.72
	PNP[26]	23.20	23.20	31.51	28.92	14.87	14.91
Ours (Margin+ρ Sampling)		23.50	23.50	31.66	29.09	14.52	14.53
Ours (Margin+ρ Sampling+GAO)		**24.96**	**24.96**	**32.32**	**29.63**	**17.59**	**17.68**

4.4.3 Performance of Different Sampling Strategies. We also include different sampling strategies for our tuple mining study using contrastive loss and margin loss. The performance is shown in Table 3. We find that given

Just Accepted

the same loss function, compared with the random method, three sampling methods all achieve improvement. In particular, while contrastive loss yields considerably worse results than margin loss with random sampling, its performance is significantly improved when using a sampling procedure similar to margin loss. Without sampling strategy (random), performance of different loss functions is very different (73.13% and 78.38%, 40.56% and 58.21%), but using the advanced sampling strategy can reduce this difference (82.07% and 82.58%, 60.23% and 65.48%). The evidence shows the importance of sampling methods. We find that the proposed sampling method achieves competitive performance when using different datasets and different loss functions. These results show that the proposed ρ Sampling improves the generalization of the model and achieves better performance.

Table 3. Comparison of R@K and NMI performance for all sampling strategies (%). MAP means MAP@1K.

Methods		ETH Food-101					Vireo Food-172					ISIA Food-500				
		R@1	R@2	R@4	MAP	NMI	R@1	R@2	R@4	MAP	NMI	R@1	R@2	R@4	MAP	NMI
Contrastive	Random [20]	63.37	74.40	83.40	17.95	45.24	73.13	80.67	86.82	20.39	49.52	40.56	50.65	60.69	3.80	35.97
	Softhard [40]	68.67	78.98	86.37	21.07	50.33	82.07	87.69	91.81	31.40	63.61	60.23	70.17	78.42	11.58	51.19
	Distance [57]	71.43	80.96	87.79	22.53	54.20	82.20	87.97	92.07	31.45	63.52	63.25	72.81	80.58	12.79	53.57
	ρ Sampling	**72.69**	**81.85**	**88.67**	**23.31**	**57.46**	**82.73**	**88.35**	**92.29**	**31.62**	**65.69**	**64.60**	**74.02**	**81.59**	**13.22**	**57.11**
Margin	Random [20]	67.30	77.59	85.39	20.56	49.32	78.38	84.76	89.50	26.87	57.96	58.21	68.43	77.14	10.97	50.51
	Softhard [40]	73.13	82.18	88.55	23.35	55.82	82.58	88.28	92.26	31.46	64.46	**65.48**	**74.94**	**82.24**	14.04	56.31
	Distance [57]	73.06	82.13	88.75	23.38	56.44	82.85	88.30	92.23	**31.79**	65.35	64.62	73.90	81.37	12.79	54.97
	ρ Sampling	**73.52**	**82.56**	**88.98**	**23.50**	**57.57**	**82.94**	**88.38**	**92.31**	31.66	**65.88**	64.67	74.13	81.65	**14.52**	**57.25**

Fig. 6 shows that the proposed ρ Sampling outperforms the baseline on all metrics. Surprisingly, we find that the dists@inter is less than the dists@intra on the baseline method, and the situation is consistent on the ρ Sampling. The abnormal phenomenon means that the method leads to potential confusion between categories and weakens the discriminative capability. Fortunately, the fusion method solves the problem and achieves the better performance on three datasets.

4.4.4 Cross-domain Evaluation. In order to further verify the generalization ability of our method, we conduct the cross-domain evaluation experiment. Specifically, as for ETH food-101 and Vireo Food-172, we train the model via one training set and evaluate the model via another test set. For example, we use the training set of ETH Food-101 for training while the test set of Vireo Food-172 for testing and vice versa. As shown in Table 4, compared with the basic setting, all methods have performance loss. However, our method still achieves the best performance on cross-domain datasets thanks to its considerations of generalization. Specifically, our method exceeds the performance of the other methods by 1-13% on R@1, 0.5-11% on MAP@1K and 1-14% on NMI.

4.4.5 Impact of Hyper-parameters. We investigate the effect of different hyper-parameter settings, *i.e.*, probability ρ in Sec. 3.2 and loss weight α in Eq. 9 on ETH Food-101 and Vireo Food-172.

Probability ρ: The influence of ρ is shown in Fig. 7 (a). We can find that best generalization corresponding to ρ equals 0.2 or 0.3, and then the performance gradually decreases as ρ continues to increase. As mentioned in Sec. 3.1, the selected ρ should ensure the existence of the boundary while maximizing some intra-class distances during training, that is, there is no confusion between the categories. Therefore, it can be explained that too many wrong labels are difficult to provide enough training information with the increasing ρ, which will confuse the whole training process and lead to the poor performance. Therefore, we need an appropriate probability ρ to make a balance between training efficiency and generalization.

Weight α: The influence of α is shown in Fig. 7 (b). We observe that the α = 0.1 results in the highest R@1. α controls the weight of the classification loss and the metric learning loss.

JUST Accepted

Table 4. Comparison of R@1 and NMI performance on domain transfer. ETH Food-101→Vireo Food-172 means that we use ETH Food-101 for training and Vireo Food-172 for testing and vice versa (%). MAP means MAP@1K.

Method	ETH Food-101→Vireo Food-172					Vireo Food-172→ETH Food-101				
	R@1	R@2	R@4	MAP	NMI	R@1	R@2	R@4	MAP	NMI
Contrastive[44]	75.06	82.34	87.94	17.64	51.32	70.45	80.36	87.54	20.81	52.74
Softmax[62]	75.09	82.40	87.96	19.64	53.53	70.21	79.69	86.45	19.92	53.68
Smooth-AP[5]	74.81	81.51	87.24	19.69	51.30	70.27	79.73	86.50	22.21	52.44
Circle[47]	63.77	73.00	80.78	10.35	42.15	57.45	68.53	76.35	13.83	43.76
Proxy Synthesis[16]	73.01	80.71	86.89	16.79	50.10	64.52	75.27	83.43	17.07	46.78
PNP[26]	75.87	83.21	88.20	19.90	53.86	70.33	79.68	86.32	22.41	52.10
Ours	**76.98**	**83.86**	**88.85**	**21.41**	**55.49**	**72.36**	**82.01**	**88.34**	**23.19**	**56.23**

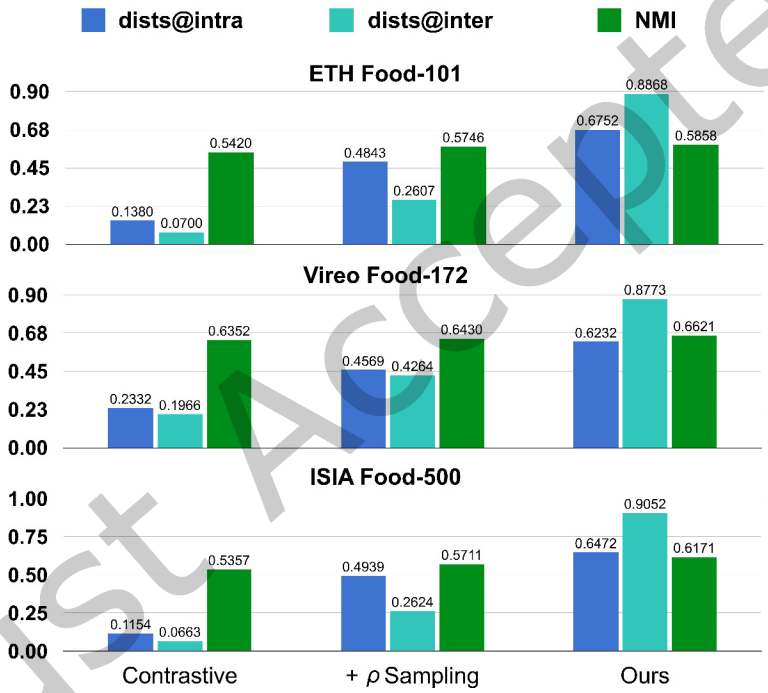

Fig. 6. The comparison of dists@intra, dists@inter and NMI on three datasets. Ours represents the combination of our methods.

Just Accepted

4.5 Qualitative Analysis

Relationship between generalization and over-fit. In Sec. 1, we point out the importance of generalization in food image retrieval and the poor generalization will lead to limited performance. In experiments, we find that popular metric learning methods, such as contrastive loss and margin loss are easy to over-fit, which are shown in Fig. 8. Although these two methods converge rapidly, the over-fitting is very obvious, especially in ISIA Food-500, which shows the limited generalization. The convergence speed of our method is slower than the

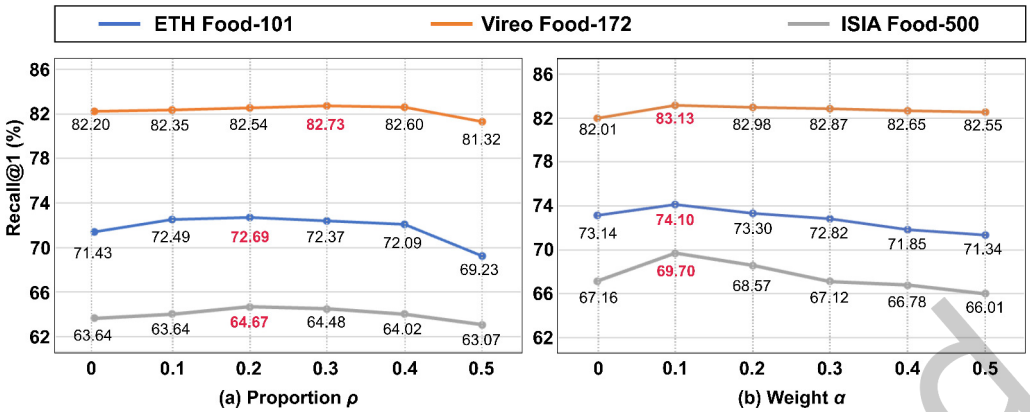

Fig. 7. Impact of different hyper-parameters on ETH Food-101, Vireo Food-172 and ISIA Food-500 (%).

other two methods, but the slow growth obtains better performance, which also shows that our method is not easy to over-fit, and our model has better generalization and superior performance.

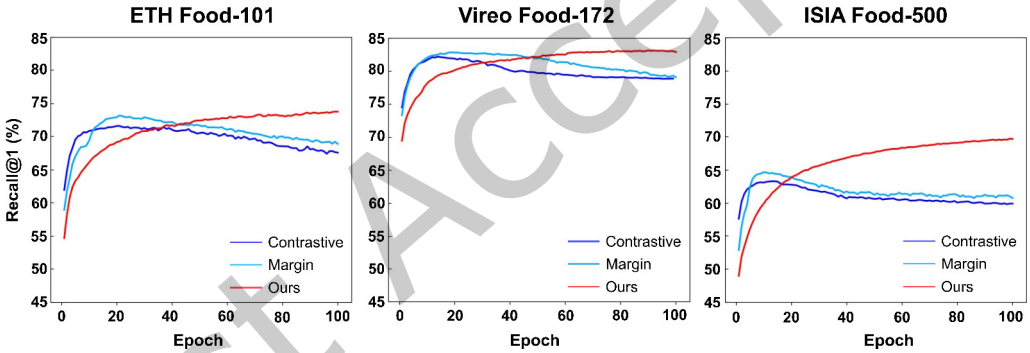

Fig. 8. The Recall@1 on three test datasets with different methods.

In this section, we qualitatively analyze the performance of the proposed method. Retrieval results on three benchmark datasets are shown in Fig. 9. We select one query on each dataset and visualize Top-10 retrieval results. We can observe that the retrieval performance achieves great improvement with our method. Significantly, as for the third query, the baseline model mistakenly recalls other food images due to their similar appearance. In contrast, our model successfully captures main features of the query, excludes hard negative instances and achieves a more favorable ranking list.

To further verify the impact of the generalization, we visualize feature embeddings optimized by the baseline, *i.e.*, contrastive loss, and our method with t-SNE [27] on three datasets. We randomly choose three thousand instances from test sets as target points. As shown in Fig. 10, compared with the baseline, our method can occupy a larger hyperspace and the discrimination of different categories is more obvious which means better generalization. Results of the visualization also correspond to larger dists@intra and dists@inter in Fig. 6.

Just Accepted

Fig. 9. Query instances and Top-10 ranked instances from the retrieval set on ETH Food-101, Vireo Food-172 and ISIA Food-500 datasets when using the baseline model (ImageNet pre-trained weights), and after using our method. Images with green border are positive instances and images with red border are negative ones for each query.

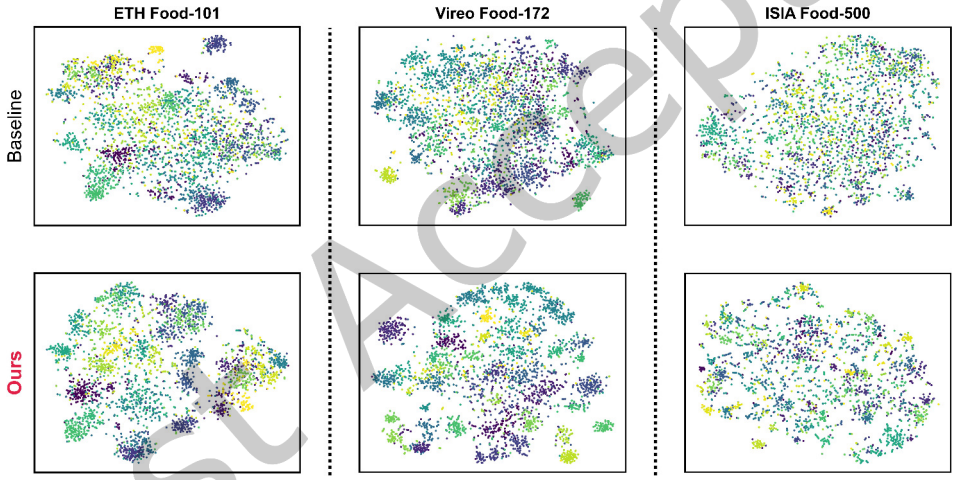

Fig. 10. Visualization of feature embeddings optimized by different methods with t-SNE [27]. Each point represents an image in test sets of three datasets. Different classes are distinguished by different colors.

Just Accepted

4.6 Performance of General Image Retrieval

Current results verify the power of our method in food image retrieval. We hope to prove that solving the problem from the generalization can also improve the performance of other tasks. In order to further verify the performance on the general image retrieval task, we experiment on three popular general datasets. **CUB200-2011** [52] is a fine-grained bird dataset, which contains 11,788 images from 200 categories. Empirically, we use the first 100 classes (5,864 images) for training and the last 100 classes (5,924 images) for testing. **CARS196** [24] has 196 car classes with 16,185 images. Training/Test sets use the first/last 98 classes (8054/8131 images). **Stanford Online Products** [46] contains 120,053 product images divided into 22,634 classes. We use the first 11,318 classes (59,551 images) for training and the last 11,316 classes (60,502 images) for testing. In SOP, unlike the other benchmarks,

most classes have few instances, leading to the significantly different data distribution compared to CUB200-2011 and CARS196.

Table 5. Comparison of R@K and NMI performance for all loss functions on three general image retrieval datasets (%). MAP means MAP@*C*.

Methods	CUB200-2011					CARS196					SOP				
	R@1	R@2	R@4	MAP	NMI	R@1	R@2	R@4	MAP	NMI	R@1	R@2	R@4	MAP	NMI
ImageNet	52.43	64.47	75.41	13.77	55.56	38.49	50.45	63.24	5.02	41.25	47.49	52.66	57.69	16.68	84.03
Contrastive	70.07	79.91	87.09	27.77	71.08	76.61	85.14	90.75	21.75	64.12	75.74	80.07	83.67	40.03	90.14
Margin	71.40	80.86	87.53	28.20	73.70	80.42	87.63	92.58	24.16	67.76	77.61	82.03	85.38	41.63	89.79
Circle	67.38	77.07	84.65	22.08	64.78	70.06	79.21	89.33	16.89	57.98	73.96	78.97	83.17	36.16	88.19
Softmax	62.42	74.34	83.31	21.69	66.70	72.44	81.87	88.72	18.79	61.40	70.18	74.89	79.13	33.32	88.21
Proxy Synthesis	62.73	73.63	83.00	21.66	66.54	70.51	80.19	87.33	17.06	59.04	68.39	73.12	77.33	31.59	87.81
Smooth-AP	71.07	80.55	87.36	28.28	71.56	75.12	83.58	89.77	21.33	62.83	76.68	81.34	85.17	40.51	89.93
PNP	71.13	80.84	**87.88**	28.27	72.30	74.49	83.30	89.23	24.17	63.41	77.53	82.07	85.82	41.58	90.12
Ours	**71.69**	**80.94**	87.71	**28.76**	**74.13**	**80.76**	**87.87**	**92.88**	**24.88**	**68.83**	**78.77**	**83.28**	**86.72**	**42.95**	**90.28**

Results show that our method still outperforms others on the general image retrieval task. Specifically, our method outperforms others by 0.5-1% on R@1, 0.5-1% on MAP@*C* and 0.2-1.1% on NMI. Specifically, we find that the performance gain on CUB200-2011 and CARS196 is lower than SOP. This is because the difference between categories of the first two datasets is not obvious. Nevertheless, our method outperforms others by 0.3% on R@1. Due to the diversity of products, SOP is similar to food datasets, so the performance improvement of our method on SOP is more obvious, which also shows that our idea of the generalization is more suitable for other image retrieval tasks with larger differences between categories.

Surprisingly, we find that proxy-based methods, which perform well on food image datasets are not effective in the general image retrieval task, *i.e.*, Softmax and Proxy Synthesis. Compared with general image retrieval datasets, the number of categories in food datasets are small, and the number of samples in each category is large. Proxy-based losses introduce the idea of classification loss and optimize samples of the same category to the center by constructing the category center. When the number of samples in a category is large, these loss functions are very effective, but pair-based losses have a lot of local optimization, which leads to convergence difficulties and affects the performance. On the contrary, general image retrieval datasets have a small number of samples per category, *i.e.*, SOP, which leads to the difficulty of constructing the center. Therefore, the performance of proxy-based losses is poor.

5 CONCLUSIONS

In this paper, we propose a generalization-oriented analysis method for food image retrieval and conduct a new benchmark on three food datasets. Focusing on the generalization problem caused by the large difference between categories of different domains in food images, we propose the ρ Sampling and GAO methods, which greatly improve the generalization of the model. The proposed method also achieves the best performance on three benchmark datasets.

Future work includes: (1) **The selection of the ρ parameter.** In this paper, we investigate the impact of the ρ parameter and conduct the evaluation on different food image datasets. However, the selection of ρ is affected by the data distribution, the number of samples in each category and the number of categories. In the future work, we hope to explore an adaptive method to obtain the optimal ρ to adapt to different datasets. (2)

Just Accepted

The generalization for food analysis. In this work, we design a generalization-oriented sampling strategy and a loss function design method to enhance the generalization of the model. In the future work, we can also design a feature extraction method to enhance the generalization. In addition, due to the in-depth research on generalization in tasks such as transfer learning, we can learn from the idea of transfer learning to design a network for food image retrieval.

6 ACKNOWLEDGEMENT

This work was supported in part by the National Nature Science Foundation of China [Grant No. 61972378, 62125207, U1936203, U19B2040] and CAAI-Huawei MindSpore Open Fund.

REFERENCES

[1] G. Albert, A. Jon, R. Jerome, and L. Diane. 2016. Deep Image Retrieval: Learning Global Representations for Image Search. In *European Conference on Computer Vision*. 241–257.
[2] G. Barlacchi, A. Abad, E. Rossinelli, and A. Moschitti. 2016. Appetitoso: A Search Engine for Restaurant Retrieval based on Dishes. In *Italian Conference on Computational Linguistics*. 46–50.
[3] A. Bellet and A. Habrard. 2015. Robustness and Generalization for Metric Learning. *Neurocomputing* 151 (2015), 259–267.
[4] L. Bossard, M. Guillaumin, and L. Van Gool. 2014. Food-101-Mining Discriminative Components with Random Forests. In *European Conference on Computer Vision*. 446–461.
[5] A. Brown, W. Xie, V. Kalogeiton, and A. Zisserman. 2020. Smooth-AP: Smoothing the Path Towards Large-Scale Image Retrieval. In *European Conference on Computer Vision*. 677–694.
[6] M. Chang, L. Guillain, H. Jung, V. Hare, J. Kim, and M. Agrawala. 2018. RecipeScape: An Interactive Tool for Analyzing Cooking Instructions at Scale. In *Proc. CHI Conference on Human Factors in Computing Systems*. 451:1–451:12.
[7] C. Chaudhary, P. Goyal, N. Goyal, and Y. Chen. 2020. Image Retrieval for Complex Queries Using Knowledge Embedding. *ACM Transactions on Multimedia Computing, Communications and Applications* 16, 1 (2020), 13:1–13:23.
[8] J. Chen and C. Ngo. 2016. Deep-based Ingredient Recognition for Cooking Recipe Retrieval. In *Proceedings of the ACM International Conference on Multimedia*. 32–41.
[9] J. Chen, C. Ngo, F. Feng, and T. Chua. 2018. Deep Understanding of Cooking Procedure for Cross-modal Recipe Retrieval. In *Proceedings of the ACM International Conference on Multimedia*. 1020–1028.
[10] G. Ciocca, P. Napoletano, and R. Schettini. 2017. Learning CNN-based Features for Retrieval of Food Images. In *International Conference on Image Analysis and Processing*. 426–434.
[11] Y. Cui, F. Zhou, Y. Lin, and S. Belongie. 2016. Fine-Grained Categorization and Dataset Bootstrapping Using Deep Metric Learning with Humans in the Loop. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1153–1162.
[12] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.
[13] T. Do, T. Hoang, D. Tan, H. Le, T. Nguyen, and N. Cheung. 2019. From Selective Deep Convolutional Features to Compact Binary Representations for Image Retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications* 15, 2 (2019), 43:1–43:22.
[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.
[15] G. M. Farinella, D Allegra, M Moltisanti, F Stanco, and S Battiato. 2016. Retrieval and Classification of Food Images. *Computers in Biology and Medicine* 77 (2016), 23–39.
[16] G. Gu, B. Ko, and H. Kim. 2021. Proxy Synthesis: Learning with Synthetic Classes for Deep Metric Learning. In *AAAI*. 1460–1468.
[17] R. Hadsell, S. Chopra, and Y. Lecun. 2006. Dimensionality Reduction by Learning an Invariant Mapping. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1735 – 1742.
[18] A. Hermans, L. Beyer, and B. Leibe. 2017. In Defense of the Triplet Loss for Person Re-Identification. *arXiv preprint arxiv: 1703.07737* (2017).
[19] S. Horiguchi, S. Amano, M. Ogawa, and K. Aizawa. 2018. Personalized Classifier for Food Image Recognition. *IEEE Trans. Multimedia* 20, 10 (2018), 2836–2848.
[20] J. Hu, J. Lu, and Y. Tan. 2014. Discriminative Deep Metric Learning for Face Verification in the Wild. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1875–1882.

Just Accepted

[21] Xin Ji, Wei Wang, Meihui Zhang, and Yang Yang. 2017. Cross-Domain Image Retrieval with Attention Modeling. In *Proceedings of the ACM International Conference on Multimedia*. 1654–1662.
[22] S. Jiang, W. Min, Y. Lyu, and L. Liu. 2020. Few-shot Food Recognition via Multi-view Representation Learning. *ACM Transactions on Multimedia Computing, Communications and Applications* 16, 3 (2020), 87:1–87:20.
[23] D. Kingma and J. Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*. 1–15.
[24] J. Krause, M. Stark, J. Deng, and F. Li. 2013. 3D Object Representations for Fine-Grained Categorization. In *IEEE International Conference on Computer Vision Workshops*. 554–561.
[25] Y. Li, H. Yao, T. Zhang, and C. Xu. 2021. Part-based Structured Representation Learning for Person Re-identification. *ACM Transactions on Multimedia Computing, Communications and Applications* 16, 4 (2021), 134:1–134:22.
[26] Z. Li, W. Min, J. Song, Y. Zhu, and S. Jiang. 2021. Rethinking Ranking-based Loss Functions: Only Penalizing Negative Instances before Positive Ones is Enough. *arXiv preprint arXiv:2102.04640* (2021).
[27] L. Maaten and G. Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605.
[28] R. Manmatha, Chao-Yuan Wu, Alexander Smola, and Philipp Krahenbuhl. 2017. Sampling Matters in Deep Embedding Learning. In *IEEE International Conference on Computer Vision*. 2859–2867.
[29] C. Micael, C. Rémi, P. David, S. Laure, T. Nicolas, and C. Matthieu. 2018. Cross-Modal Retrieval in the Cooking Context: Learning Semantic Text-image Embeddings. In *International ACM SIGIR Conference*. 35–44.
[30] T. Milbich, K. Roth, B. Brattoli, and B. Ommer. 2022. Sharing Matters for Generalization in Deep Metric Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 1 (2022), 416–427.
[31] W. Min, S. Jiang, and R. Jain. 2020. Food Recommendation: Framework, Existing Solutions, and Challenges. *IEEE Transactions on Multimedia* 22, 10 (2020), 2659–2671.
[32] W. Min, S. Jiang, L. Liu, Y. Rui, and R. Jain. 2019. A Survey on Food Computing. *Comput. Surveys* 52, 5 (2019), 36.
[33] W. Min, S. Jiang, J. Sang, H. Wang, X. Liu, and L. Herranz. 2017. Being a Supercook: Joint Food Attributes and Multi-modal Content Modeling for Recipe Retrieval and Exploration. *IEEE Transactions on Multimedia* 19, 5 (2017), 1100 – 1113.
[34] W. Min, L. Liu, Z. Wang, Z. Luo, X. Wei, X. Wei, and Jiang S. 2020. ISIA Food-500: A Dataset for Large-Scale Food Recognition via Stacked Global-Local Attention Network. In *Proceedings of the ACM International Conference on Multimedia*. 393–401.
[35] Y. Movshovitz-Attias, A. Toshev, T. Leung, S. Ioffe, and S. Singh. 2017. No Fuss Distance Metric Learning Using Proxies. In *IEEE International Conference on Computer Vision*. 360–368.
[36] Y. Movshovitz-Attias, A. Toshev, T. K. Leung, S. Ioffe, and S. Singh. 2017. No Fuss Distance Metric Learning Using Proxies. In *IEEE International Conference on Computer Vision*. 360–368.
[37] E. Ong, S. Husain, and M. Bober. 2017. Siamese Network of Deep Fisher-vector Descriptors for Image Retrieval. *arXiv preprint arXiv:1702.00338* (2017).
[38] P. Pouladzadeh and S. Shirmohammadi. 2017. Mobile Multi-Food Recognition Using Deep Learning. *ACM Transactions on Multimedia Computing, Communications and Applications* 13, 3s (2017), 36:1–36:21.
[39] Q. Qian, L. Shang, B. Sun, J. Hu, T. Tacoma, H. Li, and R. Jin. 2019. SoftTriple Loss: Deep Metric Learning Without Triplet Sampling. In *IEEE International Conference on Computer Vision*. 6449–6457.
[40] K. Roth and B Brattoli. 2019. https://github.com/Confusezius/Deep-Metric-Learning-Baselines.
[41] K. Roth, T. Milbich, S. Sinha, P. Gupta, B. Ommer, and J. Cohen. 2020. Revisiting Training Strategies and Generalization Performance in Deep Metric Learning. In *ICML*. 8242–8252.
[42] A. Salvador, N. Hynes, Y. Aytar, J. Marin, F. Ofli, I. Weber, and A. Torralba. 2017. Learning Cross-modal Embeddings for Cooking Recipes and Food Images. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3020–3028.
[43] F. Schroff, D. Kalenichenko, and J. Philbin. 2015. Facenet: A Unified Embedding for Face Recognition and Clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*. 815–823.
[44] W. Shimoda and K. Yanai. 2017. Learning Food Image Similarity for Food Image Retrieval. In *IEEE Third International Conference on Multimedia Big Data*. 165–168.
[45] K. Sohn. 2016. Improved Deep Metric Learning with Multi-class N-pair Loss Objective. In *Advances in neural information processing systems*. 1857–1865.
[46] H. Song, Y. Xiang, S. Jegelka, and S. Savarese. 2016. Deep Metric Learning via Lifted Structured Feature Embedding. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4004–4012.
[47] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei. 2020. Circle Loss: A Unified Perspective of Pair Similarity Optimization. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6398–6407.
[48] Z. Tang and J. Huang. 2022. Harmonious Multi-branch Network for Person Re-identification with Harder Triplet Loss. *ACM Transactions on Multimedia Computing, Communications and Applications* 18, 4 (2022), 98:1–98:21.
[49] N. Tishby and N. Zaslavsky. 2015. Deep Learning and the Information Bottleneck Principle. In *IEEE Information Theory Workshop*. 1–5.
[50] E. Ustinova and V. Lempitsky. 2016. Learning Deep Embeddings with Histogram Loss. In *NIPS*. 4170–4178.

Just Accepted

[51] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, and Y. Bengio. 2019. Manifold Mixup: Better Representations by Interpolating Hidden States. In *International Conference on Machine Learning*. 6438–6447.
[52] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. *The Caltech-UCSD Birds-200-2011 Dataset.* Technical Report CNS-TR-2011-001. California Institute of Technology.
[53] H. Wang, D. Sahoo, C. Liu, K. Shu, P. Achananuparp, E. Lim, and S. Hoi. 2022. Cross-Modal Food Retrieval: Learning a Joint Embedding of Food Images and Recipes With Semantic Consistency and Attention Mechanism. *IEEE Transactions on Multimedia* 24 (2022), 2515–2525.
[54] L. Wang, Q. Li, N. Li, G. Dong, and Y. Yang. 2008. Substructure Similarity Measurement in Chinese Recipes. In *Proceedings of the ACM International Conference on World Wide Web*. 979–988.
[55] W. Wang, L. Duan, H. Jiang, P. Jing, X. Song, and L. Nie. 2021. Market2Dish: Health-aware Food Recommendation. *ACM Transactions on Multimedia Computing, Communications and Applications* 17, 1 (2021), 33:1–33:19.
[56] Z. Wang, Y. Li, R. Hong, and X. Tian. 2019. Eigenvector-Based Distance Metric Learning for Image Classification and Retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications* 15, 3 (2019), 84:1–84:19.
[57] C. Wu, R. Manmatha, A. Smola, and P. Krahenbuhl. 2017. Sampling Matters in Deep Embedding Learning. In *IEEE International Conference on Computer Vision*. 2840–2848.
[58] J. Wu, J. Jiang, M. Qi, C. Chen, and Y. Liu. 2022. Improving Feature Discrimination for Object Tracking by Structural-similarity-based Metric Learning. *ACM Transactions on Multimedia Computing, Communications and Applications* 18, 4 (2022), 90:1–90:23.
[59] H. Xie, L. Yu, and Q. Li. 2011. A Hybrid Semantic Item Model for Recipe Search by Example. In *IEEE International Symposium on Multimedia*. 254–259.
[60] Q. Xu, A. Molino, J. Lin, F. Fang, V. Subbaraju, L. Li, and J. Lim. 2021. Lifelog Image Retrieval Based on Semantic Relevance Mapping. *ACM Transactions on Multimedia Computing, Communications and Applications* 17, 3 (2021), 92:1–92:18.
[61] X. Yao, D. She, H. Zhang, J. Yang, M. Cheng, and L. Wang. 2021. Adaptive Deep Metric Learning for Affective Image Retrieval and Classification. *IEEE Trans. Multimedia* 23 (2021), 1640–1653.
[62] A. Zhai and H. Wu. 2019. Classification is a Strong Baseline for Deep Metric Learning. In *British Machine Vision Conference*. 91.
[63] L. Zhang, H. Guo, K. Zhu, H. Qiao, G. Huang, S. Zhang, H. Zhang, J. Sun, and J. Wang. 2022. Hybrid Modality Metric Learning for Visible-Infrared Person Re-Identification. *ACM Transactions on Multimedia Computing, Communications and Applications* 18, 1s (2022), 25:1–25:15.
[64] S. Zhang, G. Li, W. Zhang, Q. Huang, Ti. Huang, M.Shah, and N. Sebe. 2022. Introduction to the Special Issue on Fine-Grained Visual Recognition. *ACM Transactions on Multimedia Computing, Communications and Applications* 18, 1s (2022), 24:1–24:3.
[65] B. Zhu, C. Ngo, and W. Chan. 2022. Learning From Web Recipe-Image Pairs for Food Recognition: Problem, Baselines and Performance. *IEEE Transactions on Multimedia* 24 (2022), 1175–1185.

Just Accepted