Improved Deep Unsupervised Hashing with Fine-grained Semantic Similarity Mining for Multi-Label Image Retrieval

Zeyu Ma[1], **Xiao Luo**[2*], **Yingjie Chen**[4], **Mixiao Hou**[1], **Jinxing Li**[1,3], **Minghua Deng**[2] and **Guangming Lu**[1*]

[1]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China
[2]School of Mathematical Sciences, Peking University, Beijing, China
[3]Linklogis, Shenzhen, China
[4]School of Computer Science, Peking University, Beijing, China
zeyu.ma@stu.hit.edu.cn, {xiaoluo, chenyingjie, dengmh}@pku.edu.cn, mixiaohou@163.com, {lijinxing158, luguangm}@hit.edu.cn

Abstract

In this paper, we study deep unsupervised hashing, a critical problem for approximate nearest neighbor research. Most recent methods solve this problem by semantic similarity reconstruction for guiding hashing network learning or contrastive learning of hash codes. However, in multi-label scenarios, these methods usually either generate an inaccurate similarity matrix without reflection of similarity ranking or suffer from the violation of the underlying assumption in contrastive learning, resulting in limited retrieval performance. To tackle this issue, we propose a novel method termed HAMAN, which explores semantics from a fine-grained view to enhance the ability of multi-label image retrieval. In particular, we reconstruct the pairwise similarity structure by matching fine-grained patch features generated by the pre-trained neural network, serving as reliable guidance for similarity preserving of hash codes. Moreover, a novel conditional contrastive learning on hash codes is proposed to adopt self-supervised learning in multi-label scenarios. According to extensive experiments on three multi-label datasets, the proposed method outperforms a broad range of state-of-the-art methods.

1 Introduction

Learning to hash has gained significant attention for image retrieval because of its outstanding retrieval efficiency and low storage cost [Luo *et al.*, 2022; Tu *et al.*, 2019; Wang *et al.*, 2017]. The basic principle of hashing is to compress high-dimensional data into compact binary codes while retaining their semantic similarity.

Previous hashing methods are mostly studied in the cases of supervised end-to-end training [Tu *et al.*, 2021a; Xie *et al.*, 2020; Tu *et al.*, 2021b; Wang *et al.*, 2021]. However, supervised hashing approaches are difficult to be implemented in reality due to the prohibitive cost of large-scale data annotations. Numerous deep unsupervised approaches are presented to overcome this issue and offer a cost-effective solution to practical applications, which can be mainly summarized into two categories, i.e., similarity reconstruction-based methods [Yang *et al.*, 2018; Yang *et al.*, 2019; Tu *et al.*, 2020; Shen *et al.*, 2020; Luo *et al.*, 2021a] and self-supervised learning methods [Lin *et al.*, 2016; Jang and Cho, 2021; Li *et al.*, 2021]. The first type reconstructs the binary pairwise similarity of the original data based on the pre-trained neural network, and then optimizes a hashing network for generating compact and similarity-preserving hash codes with the guidance of the reconstructed similarity structure. The second type usually enforces the hash code invariant to random augmentations. Typically, recent contrastive learning-based methods [Jang and Cho, 2021; Li *et al.*, 2021; Luo *et al.*, 2021b] propose to maximize the mutual information between the input sample and its hash code by contrasting positive pairs augmented from the same sample with negative-sampled counterparts.

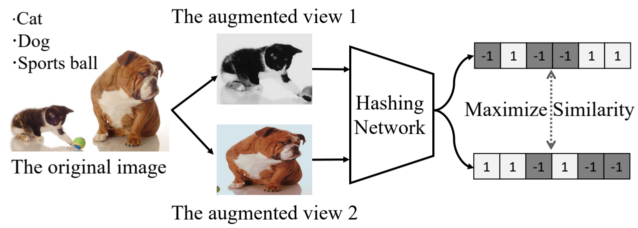

Figure 1: Motivation of our model. Random augmentations could bring in different semantics for multi-label images. Contrastive learning on hash codes maximizes the similarity of hash codes obtained from two different augmented views of the same image, even though they could have quite different semantics.

However, existing methods suffer from two limitations that can harm the quality of hash codes when it comes to more challenging multi-label image retrieval [Rodrigues *et al.*, 2020; Xie *et al.*, 2020]. On the one hand, similarity reconstruction-based methods usually define the similarity in a coarse manner, i.e., the similarities of pairwise images are

*Corresponding authors.

usually binary. Clearly, such structure cannot reconstruct the complicated similarity relationships in multi-label datasets. In particular, when two images share more labels, their similarity should be larger. Notably, such coarse similarity structure is incapable of reflecting this ranking information, far from depicting the complicated similarity structure. Owing to unreliable guidance, these methods usually accumulate a lot of errors during hash code learning. On the other hand, the underlying assumption under contrastive learning is that different augmentations of images share the same semantics, which is usually violated in multi-label scenarios. For example, as shown in Figure 1, random cropping could result in two augmented images with different semantics, which implies a false positive pair in contrastive learning, leading to a decline of performance for multi-label datasets.

To tackle the above issues, we propose a new unsupervised hashing method termed **H**ashing with fine-gr**A**ined se**M**antic simil**A**rity mi**N**ing (HAMAN) tailed for multi-label image retrieval. The core of our method is to explore semantics from a fine-grained view for improving similarity preserving learning and contrastive learning of hash codes. To explore complex similarity relationships in datasets, we split images into patches and generate patch features by the pre-trained network. Then we reconstruct the pairwise similarity structure by matching patch features of each image pair, serving as a fine-grained guidance for learning similarity-preserving hash codes. For better contrastive learning, we measure the fine-grained pairwise similarity of deep features from the augmented pair as the pseudo-label. The pseudo-label indicates whether two augmentations have the same semantics or not, serving as a condition to guide contrastive learning for discriminative hash codes. Extensive experiments on three datasets demonstrate significant and consistent improvements of HAMAN over rival baselines for multi-label image retrieval. Our main contributions are summarized as follows:

- We propose a novel deep unsupervised hashing method termed HAMAN, which mines fine-grained semantic similarity for effective multi-label image retrieval.
- We not only explore patch features for accurate similarity reconstruction, but also eliminate false positive pairs for conditional contrastive learning, producing similarity-preserving and discriminative hash codes.
- Experiments on three multi-label datasets verify that HAMAN significantly outperforms the state-of-the-art unsupervised hashing methods.

2 Related Work

Deep Unsupervised Hashing. Deep unsupervised hashing methods can be mainly classified into similarity reconstruction-based methods and self-supervised learning methods. The first category constructs the semantic structure by generating the similarity graph based on the extracted deep features. SSDH [Yang *et al.*, 2018] utilizes the Gaussian estimation to construct the semantic structure as the guide of hash code learning. DistillHash [Yang *et al.*, 2019] enhances the semantic structure by distilling image pairs and further improves the performance. MLS^{3}RDUH [Tu *et al.*, 2020] reconstructs the local semantic similarity structure based on the intrinsic manifold structure in the feature space. GLC [Luo *et al.*, 2021a] involves both global and local semantic consistency learning by clustering and similarity mining of deep features, respectively. The second type typically enforces the hash code consistent to random augmentation [Lin *et al.*, 2016; Jang and Cho, 2021; Li *et al.*, 2021]. The representative method SPQ [Jang and Cho, 2021] employs the cross quantized contrastive learning based on two different augmented views of original images. To explore the local fine-grained semantics in complicated images, our HAMAN proposes both fine-grained similarity preserving and conditional contrastive learning, producing high-quality hash codes in real-world multi-label scenarios.

Contrastive Learning. Many recent works [Wu *et al.*, 2018; Chen *et al.*, 2020; He *et al.*, 2020] indicate that unsupervised image representation learning has gained great improvement benefiting from the development of contrastive learning, which significantly reduces the gap with supervised pretraining. [Hadsell *et al.*, 2006] first attempt the representation learning by contrasting positive pairs and negative pairs. SimCLR [Chen *et al.*, 2020] adopts a simple self-supervised learning network by replacing the memory bank with elements from the same batch and achieves considerable performance on ImageNet. MoCo [He *et al.*, 2020] constructs a dynamic and consistent dictionary that preserves the candidate keys to perform contrastive learning. Considering that hash codes is a form of representation, recent researches [Jang and Cho, 2021; Li *et al.*, 2021] have brought contrastive learning into deep unsupervised hashing. However, random cropping could result in augmented images with different semantics for multi-label images, deteriorating the performance of contrastive learning. Compared with these works, we propose a novel conditional contrastive learning module, which leverages the prior to guide hash code learning.

3 The Proposed Method

To begin, the formal definition of the deep unsupervised hashing task can be explained as: $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{N}$ signifies the training set with N samples without label annotations, which is used to learn a hash function:

$$\mathcal{H} : \boldsymbol{x}_i \rightarrow \boldsymbol{b}_i \in \{-1, 1\}^l,$$

where $\boldsymbol{x}_i$ denotes the i-th input image and $\boldsymbol{b}_i$ represents the learned l-bit binary hash code. Images with similar semantic information are expected to be encoded into binary hash codes with small Hamming distances.

3.1 Framework Overview

The architecture of our hashing network $G(\cdot)$ is modified from VGG-F following previous work [Yang *et al.*, 2019; Tu *et al.*, 2020; Luo *et al.*, 2021a]. Specifically, it is constructed by substituting a fully-connected layer with l hidden units for the last fully-connected layer in VGG-F. Our hash code learning framework consists of **Fine-grained Similarity Preserving** and **Conditional Contrastive Learning**. In the first module, a feature extractor $F(\cdot)$ modified from a pre-trained VGG-F by removing the last fully-connected layer is

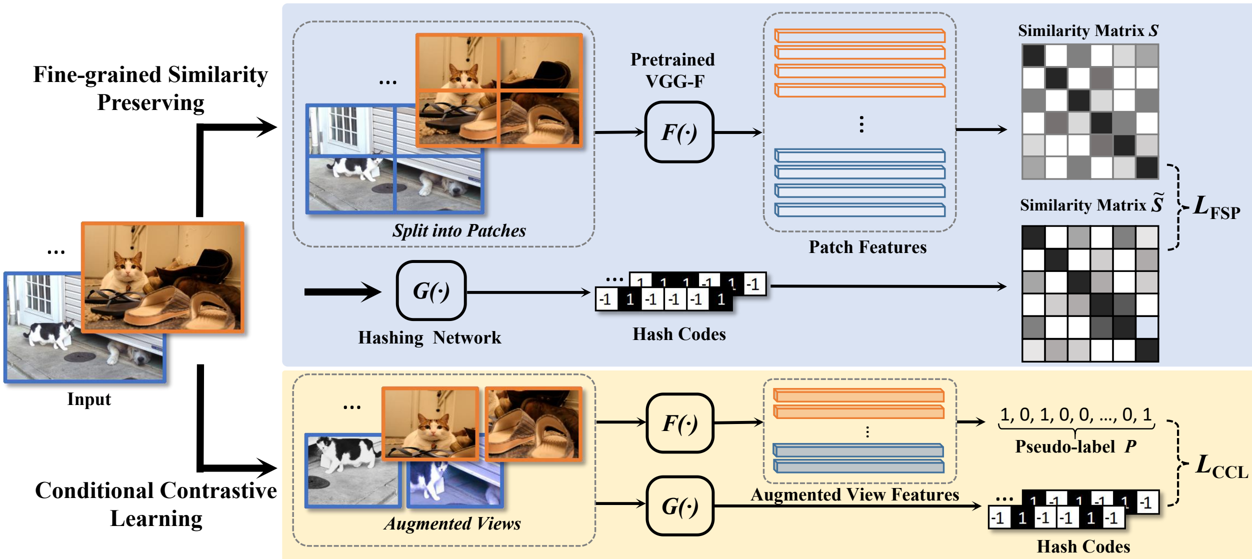

Figure 2: The framework of HAMAN. HAMAN generates a fine-grained similarity matrix, i.e., $S \in \mathbb{R}^{N\times N}$ based on the patch features of training images, providing reliable guidance for the fine-grained similarity preserving (FSP) module. Conditional contrastive learning (CCL) module is based on hash codes of augmented views and the pseudo-labels P generated from the features of augmented views.

adopted to construct a fine-grained similarity structure, which guides the hashing network to produce similarity-preserving hash codes. In the second module, we also use feature extractor $F(\cdot)$ to calculate the similarity between two views of each image, serving as the pseudo-label to facilitate conditional contrastive learning and thus improving the discriminability of hash codes. More details are illustrated in Figure 2.

3.2 Fine-grained Similarity Preserving

In this module, we first generate the fine-grained similarity structure, and train the hashing network under its guidance for generating similarity-preserving hash codes.

Fine-grained Similarity Generation

Previous methods [Tu *et al.*, 2020; Luo *et al.*, 2021a] typically use a pre-trained neural network to extract deep features for images and then calculate their pairwise cosine similarity. A binary similarity matrix can be produced by thresholding the similarity [Yang *et al.*, 2018; Yang *et al.*, 2019; Tu *et al.*, 2020]. However, For multi-label image retrieval, there exists complex ranking information. Hence, similarities of different samples should be considered in a fine-grained view.

Inspired by recent progress in Vision Transformer [Dosovitskiy *et al.*, 2021], we use local patch features to characterize the fine-granted semantics in multi-label images. Then, we match the local semantics for each image pair and summarize the comparisons for obtaining fine-granted similarity. Specifically, we split each image $\boldsymbol{x}_i \in \mathbb{R}^{H\times W\times C}$ into image patches with patch size (P, P), resulting in a sequence of patches $I_i = [I_i^1, \cdots, I_i^M] \in \mathbb{R}^{M\times P^2\times C}$. $M = HW/P^2$ is the number of patches and C is the number of channels. The feature extractor $F(\cdot)$ is used to generate the patch feature $\boldsymbol{f}_i^m = F(I_i^m)$ of each I_i^n. To obtain the similarity of image pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$, we first match patch features $\{\boldsymbol{f}_i^m\}_{m=1}^M$ of $\boldsymbol{x}_i$ with patch features $\{\boldsymbol{f}_j^n\}_{n=1}^M$ of $\boldsymbol{x}_j$. Then we select the largest similarity as local matching scores for each patch. Finally, all the matching scores are added as a summarization. The fine-grained pairwise similarity is formulated as follows,

$$S_{ij} = \frac{1}{M}\sum_{m=1}^{M} \max_n sim(\boldsymbol{f}_i^m, \boldsymbol{f}_j^n), \quad (1)$$

where $sim(\boldsymbol{u}, \boldsymbol{v})$ represents the truncated cosine similarity of two vectors, i.e., $sim(\boldsymbol{u}, \boldsymbol{v}) = max(\frac{\boldsymbol{u}^T\boldsymbol{v}}{||\boldsymbol{u}||||\boldsymbol{v}||}, 0)$. For symmetry, we take the average of S and S^T to generate the final similarity matrix. In this way, all S_{ij} are continuous values in the interval of $[0, 1]$ and if two images share more local semantics, they will obtain a larger similarity. Note that for the simplest case when $M = 1$, the fine-grained similarity is degenerated into the cosine similarity of global deep features of the whole images. When $M > 1$, our module can explore and match fine-granted semantics in images from a local view. As indicated in [Chen *et al.*, 2021], we set $M = 4$ for capturing complete local semantics as default.

Similarity Preserving Learning

For effective image retrieval, the semantic similarities between image pairs should be well preserved. In this part, we use the fine-grained pairwise similarity structure $\{S_{ij}\}_{i,j=1}^N$ to guide the training process of hashing network for producing similarity-preserving hash codes. To begin, we feed the input images into hashing network $G(\cdot)$ to produce hash codes $\{b_i\}_{i=1}^N$ and then calculate the continuous similarity matrix of hash codes as follows:

$$\tilde{S}_{ij} = \frac{\boldsymbol{b}_i^\top \boldsymbol{b}_j + L}{2L}, \quad \boldsymbol{b}_i = sign(G(x_i); \Theta)), \quad (2)$$

where Θ represents the set of hashing network parameters and the generated similarities are continuous values in $[0, 1]$. To preserve the semantic structure in a fine-granted view, we develop a mean square error loss to preserve the similarity of hash codes from the continuous pairwise similarity structure:

$$L_{FSP}\Big(\{x_i\}_{i=1}^N\Big) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\tilde{S}_{ij} - S_{ij})^2. \tag{3}$$

3.3 Conditional Contrastive Learning

In this module, we first generate pseudo-labels to decide the positive pairs and then introduce conditional contrastive learning for discriminative hash codes.

Pseudo-label Generation

Recently, self-supervised learning has shown promising results in various visual tasks and has been applied in deep unsupervised hashing [Chen *et al.*, 2020; Jang and Cho, 2021]. Their basic idea is to consider two augmented views generated from each image as positive pairs and enforce them to have similar hash codes compared with negative pairs. However, random cropping could generate views with different semantics for multi-label images. Consequently, contrastive learning of hash codes cannot achieve satisfactory performance for multi-label image retrieval. To tackle this issue, we seek to generate pseudo-labels, each of which indicates whether two augmented views constitute a positive pair with the same semantics by fine-grained similarity comparison.

Specifically, we first generate a minibatch of B sampled images and produce $2B$ randomly transformed images $\{\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)}\}_{i=1}^B$. Then we generate the pairwise pseudo-labels $\{P_i\}_{i=1}^B$ by calculating the similarities of the extracted features, which is formulated as follows:

$$P_i = \mathbf{1}_{sim(F(\boldsymbol{x}_i^{(1)}), F(\boldsymbol{x}_i^{(2)}))>\lambda}, \tag{4}$$

where λ is a pre-defined fixed similarity threshold. Only when the similarity of the image pair is above the threshold, they can be considered as a positive pair.

Conditional Contrastive Learning

Based on the pseudo-labels, we reorganize the positive pairs in the mini-batch. For the positive pair $\boldsymbol{x}_i^{(1)}$ and $\boldsymbol{x}_i^{(2)}$, the remaining $2(B-1)$ augmented views in a minibatch are considered as negative samples. Let $\boldsymbol{b}_i \star \boldsymbol{b}_j$ denote the cosine similarity of $\boldsymbol{b}_i$ and $\boldsymbol{b}_j$, and the conditional contrastive learning loss is formulated as

$$\mathcal{L}_{CCL}\Big(\{\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)}\}_{i=1}^B\Big) = -\frac{1}{2\sum_{i=1}^B P_i}\sum_{i=1}^{B} P_i\Big(\log\frac{e^{\boldsymbol{b}_i^{(1)}\star\boldsymbol{b}_i^{(2)}/\tau}}{Z_i^{(1)}} + \log\frac{e^{b_i^{(1)}\star\boldsymbol{b}_i^{(2)}/\tau}}{Z_i^{(2)}}\Big), \tag{5}$$

where $Z_i^{(a)} = \sum_{j\neq i}\left(e^{\boldsymbol{b}_i^{(a)}\star\boldsymbol{b}_j^{(1)}/\tau} + e^{\boldsymbol{b}_i^{(a)}\star\boldsymbol{b}_j^{(2)}/\tau}\right)$, $a = 1$ or 2, and τ is a temperature parameter set to 0.5 as indicated in [Chen *et al.*, 2020]. Compared with the original contrastive learning in [Jang and Cho, 2021], we introduce additional pseudo-labels as the conditions to select positive pairs for fitting the multi-label scenarios. In this way, we remove the false positive pairs in multi-label scenarios, facilitating the generation of discriminative hash codes.

Algorithm 1 Training Algorithm of HAMAN

Require: Training images $\mathcal{X} = \{x_i\}_{i=1}^N$; Code length l;
Ensure: Parameters Θ for the hashing network $G(\cdot)$;
Hash codes $\mathcal{B} = \{b_i\}_{i=1}^N$ for training images.

Split each image into four patches ;
Extract patch features of all images through $F(\cdot)$;
Construct the fine-grained pairwise similarity matrix S by Equation 1;
repeat
    Sample B images from $\mathcal{X}$ and generate $2B$ augmented images to make up a mini-batch;
    Calculate the loss by Equation 6;
    Update parameters Θ of $G(\cdot)$ by back propagation;
until convergence
Generate hash codes $\mathcal{B}$

3.4 Optimization

In summary, the loss of composite hashing network learning is formulated in the mini-batch form as

$$L = L_{CCL}(\{\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)}\}_{i=1}^H) + \eta L_{FSP}(\{\boldsymbol{x}_i\}_{i=1}^H), \tag{6}$$

where η is a balance coefficient. Notably, the parameters of the hashing network could not be updated by the standard back-propagation algorithm for the reason that the derivation of $sign(\cdot)$ is zero for any non-zero inputs and it is indifferentiable at zero. Accordingly, the $tanh(\cdot)$ is adopted to approximate the results of $sign(\cdot)$, and the approximate hash codes can be generated by using the $tanh(G(\boldsymbol{x}_i))$ to replace $\boldsymbol{b}_i$ in Equation 2 and Equation 5. In this manner, the loss functions can be optimized by the mini-batch standard stochastic gradient descent (SGD) method. For better understanding, the entire training algorithm is described in Algorithm 1.

4 Experiments

4.1 Datasets and Setup

FLICKR25K [Huiskes and Lew, 2008] contains 25,000 images with some of the 24 labels. We randomly select 2,000 images as the query set and the remaining images are used as the retrieval set, where 5,000 images are randomly selected for training. **NUS-WIDE** [Chua *et al.*, 2009] contains 269,648 images of 81 unique labels, where each image is annotated with one or more labels. Following [Jang and Cho, 2021], we use the subset with images from 21 most frequent categories. We randomly sample 100 images for each class as the query set and the rest images are used as the retrieval set, where We randomly sample 500 images for each class as the training set. **MSCOCO** [Lin *et al.*, 2014] consists of 82,783 training images and 40,504 validation images. Following [Shen *et al.*, 2020], the subset of 122,218 images from 80 categories is adopted. We randomly sample 5,000 images as the query set, and the rest images are used as the retrieval set, where 10,000 images are randomly selected for training.

Methods	FLICKR25K			NUS-WIDE			MSCOCO		
	16bits	32bits	64bits	16bits	32bits	64bits	16bits	32bits	64bits
LSH [Gionis *et al.*, 1999]	0.583	0.589	0.593	0.432	0.441	0.443	0.359	0.380	0.382
SH [Weiss *et al.*, 2009]	0.591	0.592	0.602	0.510	0.512	0.518	0.377	0.381	0.383
DeepBit [Lin *et al.*, 2016]	0.593	0.593	0.620	0.454	0.463	0.477	0.470	0.419	0.430
SGH [Dai *et al.*, 2017]	0.616	0.628	0.625	0.593	0.590	0.607	0.594	0.610	0.618
SSDH [Yang *et al.*, 2018]	0.627	0.633	0.656	0.580	0.593	0.610	0.540	0.562	0.586
DistillHash [Yang *et al.*, 2019]	0.696	0.706	0.708	0.667	0.675	0.677	0.546	0.566	0.593
CUDH [Gu *et al.*, 2019]	0.661	0.675	0.683	0.693	0.709	0.722	0.593	0.612	0.628
MLS^{3}RDUH [Tu *et al.*, 2020]	0.697	0.701	0.708	0.713	0.727	0.750	0.607	0.622	0.641
TBH [Shen *et al.*, 2020]	0.702	0.714	0.720	0.717	0.725	0.735	0.706	0.735	0.722
GLC [Luo *et al.*, 2021a]	0.758	0.772	0.777	0.759	0.772	0.783	0.715	0.723	0.731
SPQ [Jang and Cho, 2021]	0.757	0.769	0.778	0.766	0.774	0.785	-	-	-
HAMAN (Ours)	0.796	0.813	0.826	0.806	0.825	0.834	0.722	0.775	0.787

Table 1: MAP results for different methods on datasets FLICKR25K, NUS-WIDE and MSCOCO.

HAMAN is compared with a wide variety of state-of-the-art unsupervised hashing methods including two traditional shallow methods LSH [Gionis *et al.*, 1999] and SH [Weiss *et al.*, 2009], and nine deep learning methods SGH [Dai *et al.*, 2017], DeepBits [Lin *et al.*, 2016], SSDH [Yang *et al.*, 2018], DistillHash [Yang *et al.*, 2019], CUDH [Gu *et al.*, 2019], MLS^{3}RUDH [Tu *et al.*, 2020], TBH [Shen *et al.*, 2020], GLC [Luo *et al.*, 2021a] and SPQ [Jang and Cho, 2021]. For fair comparison, we use raw pixels as the input for deep learning-based methods, while 4096-dimensional feature vectors extracted by the VGG-F model, which is pre-trained on dataset ImageNet, are used for two traditional shallow methods.

The ground-truth similarity information is generated based on the ground-truth image labels. Specifically, two images are regarded as similar if they share at least one common label. We employ the Mean Average Precision (MAP), Precision-recall curve and TopN-precision curve for evaluation. For all three datasets, we adopt MAP@5000.

We implement HAMAN using PyTorch with an NVIDIA A100 80GB Tensor Core GPU. We adopt mini-batch SGD with momentum for our model training. The mini-batch size is set to 96. The learning rates for the backbone and the added fully connected layer are fixed at 0.00001 and 0.001 respectively. We resize all training images to 224×224 as inputs. We adopt five widely used techniques in the following order for data augmentation: random cropping and resizing, color jitter, random grayscale, Gaussian blur and random horizontal flip [He *et al.*, 2020]. The balance coefficient η and the similarity threshold λ are set to 1 and 0.7 as default.

4.2 Experimental Results

Table 1 shows the MAP results of HAMAN and other baselines on datasets FLICKR25K, NUS-WIDE and MSCOCO with hash code lengths of 16, 32 and 64. In addition, Figure 3 shows the Precision-recall curves and the TopN-precision curves of HAMAN and four other competitive baselines on three datasets with hash code length of 64. Accordingly, we can make the following three observations.

First, our method substantially outperforms all the competing methods with different lengths of hash codes on all three datasets. For instance, in contrast to the representative self-supervised method SPQ, HAMAN achieves an improvement of 4.4% and 4.7% for the average MAP on the dataset FLICKR25K and NUS-WIDE respectively, indicating the effectiveness of our conditional contrastive learning in positive pair selection. Second, compared with the best similarity reconstruction-based method GLC, HAMAN has 5.9%, 5.0% and 3.8% higher average MAP results on FLICKR25K, NUS-WIDE and MSCOCO, respectively. Benefiting from our fine-grained similarity exploration, HAMAN generates more accurate similarity structure of datasets and thus can im-

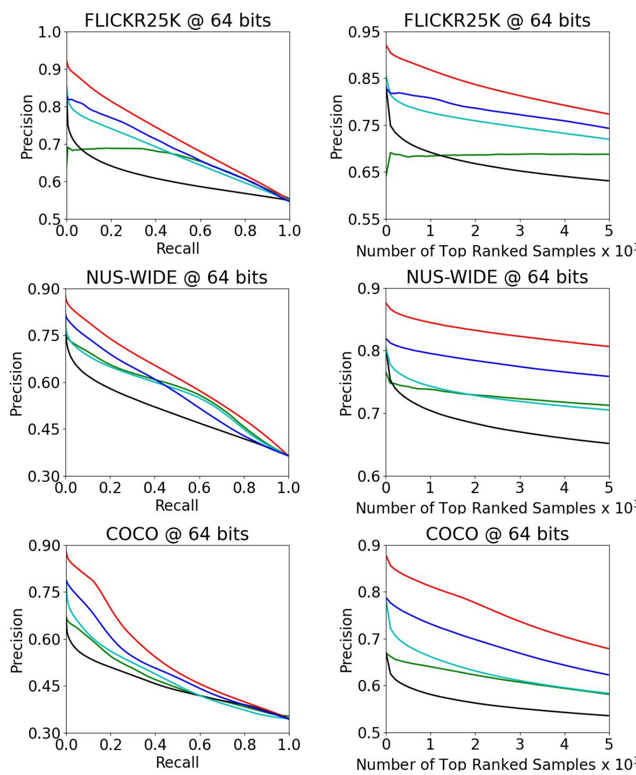

Figure 3: Precision-recall curves and Top$N$-precision curves. (— SSDH, — CUDH, — MLS^{3}RDUH, — GLC, — HAMAN)

	Components				Results		
	FSP-4	FSP-9	CL	CCL	16bits	32bits	64bits
V1	✓				0.704	0.731	0.759
V2			✓		0.652	0.667	0.687
V3				✓	0.671	0.698	0.719
V4	✓		✓		0.676	0.704	0.722
V5		✓		✓	0.711	0.766	0.771
V6	✓			✓	0.722	0.775	0.787

Table 2: Ablation analysis on MSCOCO. FSP-4, FSP-9, CL and CCL correspond to Fine-grained Similarity Preserving with $M = 4$, Fine-grained Similarity Preserving with $M = 9$, Contrastive Learning and Conditional Contrastive Learning, respectively.

prove the performance for multi-label image retrieval. Third, it can be clearly found that HAMAN achieves the best performance regarding the Precision-recall curves by comparing with several competing baselines. Fourth, as indicated by the TopN-precision curves, HAMAN outperforms other compared methods by a large margin, which demonstrates that our method can realize more effective image retrieval.

Ablation Study. To investigate the influence of different components of the proposed method, we configure several variants of HAMAN and conduct experiments on the dataset MSCOCO for comparison: (1) V1 only contains the fine-grained similarity preserving module with M set to 4 by default. (2) V2 only uses the standard contrastive learning loss [Chen *et al.*, 2020] on hash codes generated from two random augmented views of the same images. (3) V3 employs our conditional contrastive learning module with the pseudo-labels. (4) V4 involves both the fine-grained similarity preserving module and the basic contrastive learning on hash code. (5) The difference between V5 and V6 (our full model) is that V5 makes use of the fine-grained similarity preserving module with $M = 9$ and V6 uses the fine-grained similarity preserving module with $M = 4$. The results are shown in Table 2. We can have similar observations on the other datasets. To begin, we find that V3 outperforms V2 by a large margin and the performance of V4 decreases greatly in contrast to V1, which indicates that basic contrastive learning could bring in a negative effect on multi-label image retrieval and meanwhile demonstrates the superiority of our conditional contrastive learning strategy. In addition, V6 achieves significant improvements over V1 and V3, revealing that both fine-grained similarity preserving module and conditional contrastive learning module can contribute to

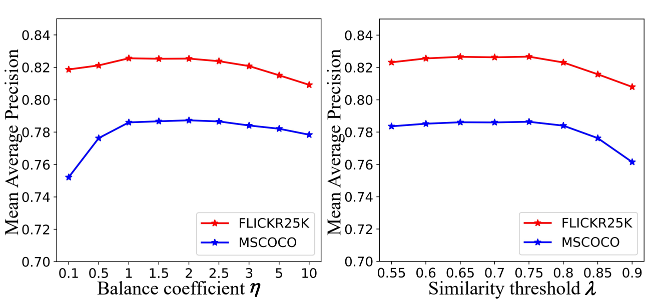

Figure 4: Sensitivity analysis of λ and η with 64-bit hash codes.

Figure 5: Examples of the top 10 returned images and Precision@10. Images in green/red boxes are correct/false results.

the improvement of our model and are appropriate for multi-label hash lookup retrieval. Lastly, V5 differs from V6 with respect to the number of split patches of training images. It can be inferred from the results that $M = 4$ is more beneficial to the fine-grained similarity generation. The potential reason is that too small patches make it hard to capture complete semantics. Hence, M is set to 4 in our model as default.

Parameter Sensitivity. We study the effect of the balance coefficient η and the similarity threshold λ through experiments on datasets FLICKR25K and MSCOCO with 64-bit hash codes. Referring to the left column of Figure 4, the performance of HAMAN is not sensitive to the value of η in the range of $[1, 2.5]$. The similarity threshold λ plays an important role in selecting positive pairs for conditional contrastive learning, as shown in the right column of Figure 4, our method can achieve the considerable performances with λ ranging from 0.55 to 0.75. Hense, η and λ are set as 1 and 0.7 for other experiments as default, respectively.

Visualization. We show the top 10 returned images of our method and GLC on FLICKR25K based on 64-bit hash codes in Figure 5. Benefiting from the fine-grained similarity mining and conditional contrastive learning in our method, HAMAN can successfully retrieval relevant images from the aspect of multiple semantics for multi-label image retrieval.

5 Conclusion

In this paper, we propose a novel deep unsupervised hashing method termed HAMAN for multi-label image retrieval. Our HAMAN consists of a fine-grained similarity preserving module and a conditional contrastive learning module, which explore the semantics of images from a fine-granted view. Experiments on three well-known benchmarks validate the efficacy of our approach. In future work, we expect to further extend our HAMAN to a broader range of applications such as cross-modal hashing and semi-supervised hashing.

Acknowledgements

This work was supported in part by the NSFC fund 62176077, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2019Bl515120055, in part by the Shenzhen Key Technical Project under Grant 2020N046, in part by the Shenzhen Fundamental Research Fund under Grant JCYJ20210324132210025, in part by Shenzhen Science and Technology Program RCBS20200714114910193 and the NSFC fund 61906162, and in part by the Medical Biometrics Perception and Analysis Engineering Laboratory, Shenzhen, China.

References

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[Chen *et al.*, 2021] Pengguang Chen, Shu Liu, and Jiaya Jia. Jigsaw clustering for unsupervised visual representation learning. In *CVPR*, 2021.

[Chua *et al.*, 2009] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *ACM CIVR*, 2009.

[Dai *et al.*, 2017] Bo Dai, Ruiqi Guo, Sanjiv Kumar, Niao He, and Le Song. Stochastic generative hashing. In *ICML*, 2017.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[Gionis *et al.*, 1999] Aristides Gionis, Piotr Indyk, Rajeev Motwani, et al. Similarity search in high dimensions via hashing. In *VLDB*, 1999.

[Gu *et al.*, 2019] Yifan Gu, Shidong Wang, Haofeng Zhang, Yazhou Yao, and Li Liu. Clustering-driven unsupervised deep hashing for image retrieval. *Neurocomputing*, 368:114–123, 2019.

[Hadsell *et al.*, 2006] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006.

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[Huiskes and Lew, 2008] Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *ACMMM*, 2008.

[Jang and Cho, 2021] Young Kyun Jang and Nam Ik Cho. Self-supervised product quantization for deep unsupervised image retrieval. In *ICCV*, 2021.

[Li *et al.*, 2021] Shuyan Li, Xiu Li, Jiwen Lu, and Jie Zhou. Self-supervised video hashing via bidirectional transformers. In *CVPR*, 2021.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[Lin *et al.*, 2016] Kevin Lin, Jiwen Lu, Chu-Song Chen, and Jie Zhou. Learning compact binary descriptors with unsupervised deep neural networks. In *CVPR*, 2016.

[Luo *et al.*, 2021a] Xiao Luo, Daqing Wu, Chong Chen, Jinwen Ma, and Minghua Deng. Deep unsupervised hashing by global and local consistency. In *ICME*, 2021.

[Luo *et al.*, 2021b] Xiao Luo, Daqing Wu, Zeyu Ma, Chong Chen, Minghua Deng, Jinwen Ma, Zhongming Jin, Jianqiang Huang, and Xian-Sheng Hua. Cimon: Towards high-quality hash codes. In *IJCAI*, 2021.

[Luo *et al.*, 2022] Xiao Luo, Haixin Wang, Chong Chen, Huasong Zhong, Hao Zhang, Minghua Deng, Jianqiang Huang, and Xian-Sheng Hua. A survey on deep hashing methods. *ACM Transactions on Knowledge Discovery from Data*, 2022.

[Rodrigues *et al.*, 2020] Josiane Rodrigues, Marco Cristo, and Juan G Colonna. Deep hashing for multi-label image retrieval: a survey. *Artificial Intelligence Review*, 53(7):5261–5307, 2020.

[Shen *et al.*, 2020] Yuming Shen, Jie Qin, Jiaxin Chen, Mengyang Yu, Li Liu, Fan Zhu, Fumin Shen, and Ling Shao. Auto-encoding twin-bottleneck hashing. In *CVPR*, 2020.

[Tu *et al.*, 2019] Rong-Cheng Tu, Xian-Ling Mao, Bo-Si Feng, and Shu-Ying Yu. Object detection based deep unsupervised hashing. In *IJCAI*, 2019.

[Tu *et al.*, 2020] Rong-Cheng Tu, Xian-Ling Mao, and Wei Wei. Mls3rduh: Deep unsupervised hashing via manifold based local semantic similarity structure reconstructing. In *IJCAI*, 2020.

[Tu *et al.*, 2021a] Rong-Cheng Tu, Xian-Ling Mao, Jia-Nan Guo, Wei Wei, and Heyan Huang. Partial-softmax loss based deep hashing. In *WWW*, 2021.

[Tu *et al.*, 2021b] Rong-Cheng Tu, Xian-Ling Mao, Cihang Kong, Zihang Shao, Ze-Lin Li, Wei Wei, and Heyan Huang. Weighted gaussian loss based hamming hashing. In *ACMMM*, 2021.

[Wang *et al.*, 2017] Jingdong Wang, Ting Zhang, Nicu Sebe, Heng Tao Shen, et al. A survey on learning to hash. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):769–790, 2017.

[Wang *et al.*, 2021] Yimu Wang, Bo Xue, Quan Cheng, Yuhui Chen, and Lijun Zhang. Deep unified cross-modality hashing by pairwise data alignment. In *IJCAI*, 2021.

[Weiss *et al.*, 2009] Yair Weiss, Antonio Torralba, and Robert Fergus. Spectral hashing. In *NeurIPS*, 2009.

[Wu *et al.*, 2018] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018.

[Xie *et al.*, 2020] Yanzhao Xie, Yu Liu, Yangtao Wang, Lianli Gao, Peng Wang, and Ke Zhou. Label-attended hashing for multi-label image retrieval. In *IJCAI*, 2020.

[Yang *et al.*, 2018] Erkun Yang, Cheng Deng, Tongliang Liu, Wei Liu, and Dacheng Tao. Semantic structure-based unsupervised deep hashing. In *IJCAI*, 2018.

[Yang *et al.*, 2019] Erkun Yang, Tongliang Liu, Cheng Deng, Wei Liu, and Dacheng Tao. Distillhash: Unsupervised deep hashing by distilling data pairs. In *CVPR*, 2019.