Hierarchical Quantization Consistency for Fully Unsupervised Image Retrieval

Guile Wu
guile.wu@outlook.com
Chao Zhang
chao.zhang@toshiba.eu
Stephan Liwicki
stephan.liwicki@toshiba.eu

Cambridge Research Lab
Toshiba Europe Ltd
Cambridge, UK

Abstract

Unsupervised image retrieval aims to learn an efficient retrieval system without expensive data annotations. Typical methods rely on handcrafted feature descriptors or pre-trained feature extractors. Recent advances propose deep fully unsupervised image retrieval aiming at training a deep model from scratch to jointly optimize visual features and quantization codes with minimal human supervision. Thus approach mainly focuses on instance contrastive learning without using semantic information. However, a fundamental problem of contrastive learning is mitigating the effects of false negatives. To this end, we exploit sub-quantized representations to extract fine-grained semantics for self-supervised learning. To further regularize the instance contrastive learning for quantization, we also leverage consistency regularization to reflect the similarities between the query sample and negative samples. Specifically, we propose a novel hierarchical consistent quantization approach to deep fully unsupervised image retrieval, which consists of part consistent quantization and global consistent quantization. With a unified learning objective, our approach exploits richer self-supervision cues to facilitate model learning. Extensive experiments on three benchmark datasets show the superiority of our approach over the state-of-the-art methods.

1 Introduction

Image retrieval is a fundamental task in computer vision, aiming to find images that are visually similar to a given query image from a large database. To reduce computational cost and improve storage efficiency, approximate nearest neighbor search [16] has been widely used, where hashing [9, 32, 39, 42] and product quantization [17, 18, 19, 34] are two most representative directions. Hashing methods map real-value embeddings to binary codes for efficient retrieval, while product quantization divides real-value data space into disjoint partitions to quantize embeddings for efficient retrieval. In the past decade, the unprecedented success of deep learning in computer vision has brought a great breakthrough to deep supervised hashing [3, 23, 42] and deep supervised quantization [2, 19, 41] based image retrieval. Although deep supervised image retrieval methods have shown outstanding performance, the

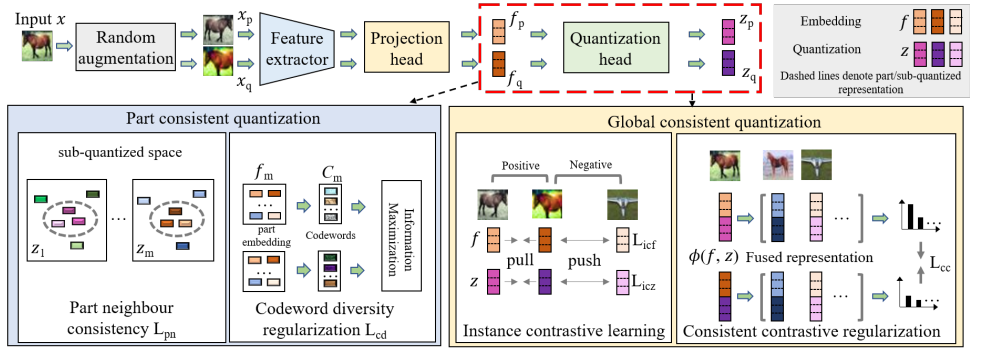


Figure 1: An overview of the proposed Self-Supervised Consistent Quantization (SSCQ) approach to deep fully unsupervised image retrieval. Part consistent quantization discovers part neighbor affinity as self-supervision, while global consistent quantization learns instance affinity as self-supervision, which together are formulated into a unified learning objective for model optimization.

reliance of expensive label annotations of training data hinders their applications in label-limited scenarios.

On the other hand, unsupervised image retrieval is capable of learning an efficient retrieval system without using labeled training data. Traditional unsupervised image retrieval methods [1, 12, 25] utilize handcrafted descriptors to extract embeddings of input images and adopt unsupervised hashing or quantization approaches to efficient retrieval. Recent deep unsupervised image retrieval methods [24, 34, 39] resort to ImageNet [30] pre-trained deep neural networks for feature extraction and incorporate deep hashing or deep quantization into a deep model for optimization. However, these methods either rely on human supervision for devising effective handcrafted feature descriptors or supervised pre-trained ImageNet backbone networks. To minimize human supervision, deep fully unsupervised image retrieval is recently proposed in [17], which aims to train a deep model from scratch to jointly optimize visual features and codes for efficient image retrieval. Despite the self-supervised product quantization approach introduced in [17] which has shown promising performance, it only applies instance contrastive loss and makes the assumption that different images are "negative" samples. However, false negatives will result in sub-optimal performance due to discarding common semantic content [15].

To reduce the effect of false negatives, we introduce a novel Self-Supervised Consistent Quantization (SSCQ) approach to deep fully unsupervised image retrieval. An overview of the proposed approach is depicted in Fig. 1. Our motivation for global consistent quantization is simple: the query sample and its positive sample should have consistent similarities to negative samples. Unlike [35], we explore this idea to tackle unsupervised image retrieval task by fusing the embedding and quantized representations. Meanwhile, since it is inevitable that the quantization process will lose useful information of embedding representations, we present a simple but effective solution to simultaneously optimize visual features and codes so as to make up for the loss. We term this process *global consistent quantization*.

It has been shown that neighbors in embedding space usually share semantic information [13, 37]. Since the quantization process in product quantization is akin to clustering, our hypothesis is that neighbors in sub-quantized space also share similar semantic information. This has a natural link to sub-quantized representations which share codewords from

a specific codebook. Therefore, we explicitly enforce sub-quantized representations to be semantically consistent. We term this process *part consistent quantization*.

We formulate both global and part consistent quantization into a unified learning objective to explore richer self-supervision for deep fully unsupervised image retrieval. To evaluate the effectiveness of the proposed approach, we conduct extensive experiments on three benchmark datasets, namely CIFAR-10 [21], NUS-WIDE [6] and FLICKR25K [14]. In summary, our **contributions** are:

1. We propose a novel hierarchical consistent quantization approach and achieve the state-of-the-art performance for deep fully unsupervised image retrieval.

2. At global level, we improve retrieval performance by exploiting contrastive consistency from "negative" instances using fused embedding and quantized representations.

3. At part level, we employ neighbor semantic consistency learning in a self-supervised way. This helps reduce the adverse effect of false negatives in contrastive learning and learn more discriminative representations for image retrieval.

2 Related Work

Self-Supervised Representation Learning. In recent years, self-supervised/unsupervised representation learning has made great progress. The common practice is devising different pretext tasks, such as predicting image rotation [20] and solving jigsaw puzzles [27], to generate self-supervision information to facilitate unsupervised representation learning. Recently, contrastive learning [5, 11, 35, 38] has become one of the most popular and powerful paradigms for unsupervised representation learning. It usually applies strong random augmentation to each input image to generate positive counterparts and employs a contrastive loss to pull the positives closer and push the negatives apart, where different instances are considered as negatives. In deep fully unsupervised image retrieval, [17] introduces a contrastive quantization framework which shows promising performance compared with conventional unsupervised image retrieval methods even without supervised pre-training. Our approach introduces self-supervised consistent quantization to discover underlying neighbor semantic structure information from sub-quantized representations and to learn affinity between instance so as to facilitate model learning in deep fully unsupervised image retrieval.

Unsupervised Image Retrieval. Most traditional image retrieval methods are originally designed for unsupervised learning scenarios where no labeled training data are used for model learning. The common practice usually consists of two disjoint steps. The first step is to use handcrafted descriptors, such as GIST [28] and SIFT [26], to extract features of input images, while the second step is to employ binary hashing [4, 31, 36] or product quantization [1, 9, 18] to transform embedding space into Hamming space or a Cartesian product of subspaces for efficient image retrieval.

In the past decade, deep learning based methods have dominated the field of image retrieval. Supervised deep hashing [23, 42] or deep product quantization [19, 41] based image retrieval methods have shown notably better performance than the traditional counterparts. However, these supervised methods are limited by the availability of labeled data for model training. To resolve this problem, researchers have resorted to deep unsupervised learning for image retrieval. One of the most popular direction is deep unsupervised hashing [24, 39].

An ImageNet [30] pre-trained deep neural network is usually used as the feature encoder and then hashing layers are inserted into the model to learn discriminative binary codes using unlabeled data.

Despite outstanding performance has been achieved, deep unsupervised hashing relies heavily on the pre-trained feature encoder and can only learn restricted binary codes, limiting its ability for distinguishing visually similar but semantic dissimilar data. Since product quantization is capable of learning continuous representations for efficient image retrieval, deep unsupervised product quantization [17, 34] is recently introduced for unsupervised image retrieval. In [34], soft quantization [41] is combined with contrastive learning [5] and code memory to learn a retrieval system in an unsupervised manner. However, [34] still requires using a pre-trained model as the feature extractor and most layers are not optimized during model training. To minimize human supervision, deep fully unsupervised image retrieval is introduced in [17], which aims to train a deep model from scratch for efficient image retrieval. In [17], a cross-quantized contrastive learning framework is proposed to jointly optimize visual features and codes for unsupervised image retrieval.

Our work belongs to deep unsupervised product quantization and focuses on deep fully unsupervised image retrieval without data label annotation nor supervised pre-trained backbone models. We propose a novel self-supervised consistent quantization approach to discover richer self-supervision to facilitate model optimization. We devise part consistent quantization to discover underlying neighbor semantic structure information and global consistent quantization to learn affinity between instances. This differs from [17] that only uses contrastive learning to optimize cross-quantized representations or [34] that requires a pre-trained model as the feature encoder as well as an additional code memory.

3 Methodology

Problem Statement. In this work, we target *deep fully unsupervised image retrieval* [17], where neither labeled training data nor pre-trained models are available. Given an unlabeled training set $\mathcal{X}=\{x^i\}_{i=1}^N$ with N samples, our task is to learn a model to encode x^i into a L-bit code $B^i=\{b_j^i\}_{j=1}^L$, where $b_j^i \in \{0,1\}$, for efficient image retrieval. During inference, the similarity between query and database samples are measured based on the learned representations and codes so as to realize efficient retrieval.

3.1 Approach Overview

An overview of the proposed Self-Supervised Consistent Quantization (SSCQ) approach is depicted in Fig. 1. In each training mini-batch $\{x^i\}_{i=1}^{N_b}$, we apply strong random data augmentation [5] on each input sample to generate two augmented views x_p^i and x_q^i, so we have $2N_b$ augmented samples in each mini-batch. Then, we extract D-dimensional embedding representations f_p^i and f_q^i of augmented inputs x_p^i and x_q^i and further quantize the embeddings into D-dimensional quantized representation z_p^i and z_q^i.

To construct the model training objective, we design hierarchical quantization to explicitly consider the self-supervised information at global and part level. Note, the meaning of *consistent* refers to two levels: at part level, the semantic consistency of sub-quantized representation is used to alleviate the adverse effect of potential false negatives; at instance level, the affinity similarity consistency to negative samples is employed to learn stable embeddings under random augmentations.

3.2 Hierarchical Self-Supervised Consistent Quantization

A Baseline with Contrastive Quantization. We build our method on a contrastive quantization baseline model introduced in [17] for fully unsupervised image retrieval. As shown in Fig. 1, with each augmented input sample x, a feature extractor is used to extract feature of x and a projection head is employed to map the learned feature into a D-dimensional embedding representation f. Then, a quantization head is used to reconstruct f into a quantized representation $z \in \mathbb{R}^D$.

Suppose there are M codebooks $\{C_m\}_{m=1}^M$ in the quantization head and each codebook is composed of K codewords $C_m=\{c_{m,k}\}_{k=1}^K$, where $c_{m,k}\in\mathbb{R}^{D/M}$. Following product quantization [18, 41], f is divided into M disjoint sub-embedding representation $f_m\in\mathbb{R}^{D/M}$ and the codewords in the m-th codebook are used to reconstruct the sub-embedding representation f_m. Therefore, the embedding space is divided into a Cartesian product of M subspaces $\{C_1\times C_2\times ...\times C_M\}$, and codewords in the m-th codebook are considered as distinct cluster centroids of the m-th sub-embedding representations of all samples. This allows to assign visually similar sub-embedding representations to the same codeword for efficient similarity measurement. To train the feature extractor, the projection head and the quantization head in an end-to-end manner, soft quantization [41] is employed for training, so the sub-quantized representation z_m of m-th codebook is defined as:

$$z_m = \sum_{k=1}^{K} \frac{\exp(d(f_m, c_{m,k})/\tau_{sq})}{\sum_{j=1}^{K}\exp(d(f_m, c_{m,j})/\tau_{sq})} c_{m,k}, \tag{1}$$

where $d(f_m, c_{m,k}) = -\|f_m - c_{m,k}\|_2^2$ is the squared Euclidean distance, and τ_{sq} is a temperature parameter, z is the concatenation of z_m.

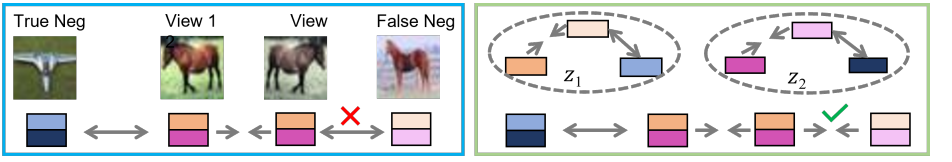
To optimize the model, an instance contrastive learning loss L_{icz} [5] is minimized to pull z closer to its positive and push away from its negatives, as:

$$\mathcal{L}_{icz} = -\log \frac{\exp(s(z, z^+)/\tau_{ic})}{\sum_{j=1}^{2N_b} 1_{[z_j \neq z]} \exp(s(z, z_j)/\tau_{ic})}, \tag{2}$$

where z and z^+ are the query and positive sample, $s(z_i, z_j) = z^T z_j/(\|z\|\|z_j\|)$ is cosine similarity between representations z and z_j, $1_{[z_j\neq z]}$ denotes an indicator function, and τ_{ic} is a temperature parameter.

With the baseline model, SPQ [17] proposed a cross quantized contrastive learning loss to jointly learn embedding representation f and the quantized representation z. This could effectively reduce the discrepancy caused by quantization and improve the performance. However, one major limitation of this work is that only instance (global) level representation is considered in the loss function. This could lead to sub-optimal performance for product quantization based method, since the potential false negatives could discard common semantic information. To overcome this limitation, we propose part consistent loss to promote semantic consistency for sub-quantized representations. Furthermore, we leverage consistency regularization using fused global representation for stable feature learning. These two novel changes enable us to explore richer self-supervision signals for unsupervised image retrieval.

Part Semantic Consistent Quantization. The quantization process in the quantization head is akin to clustering. It learns distinct codewords as cluster centroids with learnable



(a) Instance contrastive loss makes mistakes for false negative (b) Part semantic loss reduces false negative error

Figure 2: Given two views of the query instance of a *horse*, we illustrate the benefit of using part semantic loss with a true negative (*plane*) and a false negative (*another horse*). In (a), the instance contrastive loss with false negatives leads to sub-optimal feature representation. In (b), part embeddings of the anchor instance could be pulled closer to those from the *other horse*, thereby fixing the error caused by false negative in (a).

parameters for end-to-end model training. As such, each sub-quantized representation z_m inherently encodes semantic information, which can be exploited as auxiliary self-supervision cues to facilitate model learning. We use part and sub-quantized representation interchangeably. The idea is to enhance the discriminativeness of part representation, and at the same time to produce compact semantic distributions for each part representation. We achieve this by mining the K-nearest neighbors from the mini-batch. This process is illustrated in Fig.2. Specifically, we use a part neighbor semantic consistency loss $\mathcal{L}_{pn}$ to pull each part representation z_m closer to its similar neighbors and push z_m away from dissimilar ones, as:

$$\mathcal{L}_{pn} = -\frac{1}{M}\sum_{m=1}^{M}\log\frac{\sum_{n=1}^{N_k}\exp(s(z_m, z^-_{m,n})/\tau_{pn})}{\sum_{j=1}^{2Nb-2}\exp(s(z_m, z^-_{m,j})/\tau_{pn})}, \tag{3}$$

where $\{z^-_{m,n}\}_{n=1}^{N_k}$ are the top N_k subspace neighbors of z_m which are obtained by computing the similarity between z_m and its negative quantized representations $z^-_{m,j}$. The positive instance normaly is highly similar to the query and is skipped to not overwhelm other terms. Different from reducing the number of codewords, this loss promotes the compactness of semantic structure in a soft way. Thus, it does not suffer from reduced diversity of retrieval results or reduced performance.

Global Affinity Consistent Quantization. At instance level, the contrastive learning loss is effective to capture semantic information by treating two augmented views as a positive pair. Recent advances in unsupervised representation learning [22, 35] explore the affinity between query (positive) and negative instances in contrastive learning for better generalization. We introduce an affinity consistent contrastive loss $\mathcal{L}_{cc}$ to the global quantization. This loss could be applied on either quantized representations or embedding representations. We experimentally found that using the fused representations gives the best results. Concretely, we first combine f and z to learn the fused representation $\Phi(f,z)$ where $\Phi(\cdot,\cdot)$ is the fusion operation (e.g., concatenation or sum fusion). Then, we compute the similarity $Q(i)$ between $\Phi(f,z)$ and its negatives $\{\Phi(f^-,z^-)_i\}$. Similarly, the similarity $P(i)$ between $\Phi(f^+,z^+)$ and the same negatives, as:

$$\begin{aligned} Q(i) &= \frac{\exp(s(\Phi(f,z),\Phi(f^-,z^-)_i)/\tau_{cc})}{\sum_{j=1}^{2N_b-2}\exp(s(\Phi(f,z),\Phi(f^-,z^-)_j)/\tau_{cc})}, \\ P(i) &= \frac{\exp(s(\Phi(f^+,z^+),\Phi(f^-,z^-)_i)/\tau_{cc})}{\sum_{j=1}^{2N_b-2}\exp(s(\Phi(f^+,z^+),\Phi(f^-,z^-)_j)/\tau_{cc})}, \end{aligned} \tag{4}$$

Thus, contrastive consistency loss $\mathcal{L}_{cc}$ is defined using the symmetric Kullback-Leibler Divergence D_{KL}, as:

$$\mathcal{L}_{cc} = \frac{1}{2}(D_{KL}(P\|Q) + D_{KL}(Q\|P)). \tag{5}$$

Summary. In training, a unified learning objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{icz} + \mathcal{L}_{icf} + \lambda_{pn}\mathcal{L}_{pn} + \lambda_{cd}\mathcal{L}_{cd} + \lambda_{cc}\mathcal{L}_{cc}, \tag{6}$$

where λ_{pn}, λ_{cd} and λ_{cc} are weighting parameters. Among these losses, our main contributions are $\mathcal{L}_{pn}$ for part semantic consistency and $\mathcal{L}_{cc}$ for global affinity consistency. Here, $\mathcal{L}_{pn}$ differs from existing unsupervised neighbor discovery methods [13, 37, 40] in that $\mathcal{L}_{pn}$ mines neighbor affinity in the sub-quantized representations z_m, instead of progressively exploring anchored neighbors [13] or maintaining a patch/tracklet memory bank [37, 40]. Unlike [35], our $\mathcal{L}_{cc}$ uses fused representation and is more flexible. Note, $\mathcal{L}_{icf}$ is similar to $\mathcal{L}_{icz}$ except that z is replaced with f, and $\mathcal{L}_{cd}$ is a plug-and-play term for codeword diversity. We refer to the *supp. mat.* for the details of inference stage and the summary of the proposed implementation.

4 Experiments

4.1 Dataset and Evaluation Protocol

Datasets. To evaluate the proposed self-supervised consistent quantization approach for fully unsupervised image retrieval, we conduct extensive experiments on three datasets, namely CIFAR-10 [21], NUS-WIDE [6] and FLICKR25K [14].

CIFAR-10 consists of 60,000 images of 10 classes, where each class has 5,000 images for training and 1,000 images for testing. We use 1,000 images per class as the query set, while the remaining images are used as the training set and the retrieval database.

NUS-WIDE is a multi-label large-scale dataset with around 270,000 images of 81 categories. We select images of the 21 most frequent categories for evaluation, where 100 images per categories are selected to form 21,000 images as the query set while the remaining images form the training set and the retrieval database.

FLICKR25K is a relatively small dataset with 25,000 images of 24 categories. We randomly select 2,000 images as the query set while the remaining images are used as the training set and the retrieval database. On the multi-label NUS-WIDE and FLICKR25K, if a query image and a database image share at least one label, then they are defined as the true match [17, 24].

Evaluation Metrics. Following [17, 24, 32, 34], we mainly employ mean Average Precision (mAP, %) as the evaluation metric. We use mAP@1000 for CIFAR-10 and mAP@5000 for NUS-WIDE and FLICKR25K, and report image retrieval results with $\{16, 32, 64\}$ bits codes. Besides, we also report Precision-Recall curves (PR) and Precision curves with top-1000 returned samples (P@1000) at 32 bits codes. For implementation details and hyper-parameters, please refer to the *supp. mat.*

Dataset	Method	16 bits	32 bits	64 bits
CIFAR-10	SGH [2]	43.5	43.7	43.3
	HashGAN [8]	44.7	46.3	48.1
	BinGAN [43]	47.6	51.2	52.0
	SPQ [17]	76.8	79.3	81.2
	SSCQ (ours)	**78.3**	**81.3**	**82.9**
NUS-WIDE	SGH [2]	59.3	59.0	60.7
	HashGAN [8]	68.4	70.6	71.7
	BinGAN [43]	65.4	70.9	71.3
	SPQ† [17]	75.7	79.4	80.2
	SSCQ (ours)	**78.7**	**79.9**	**80.8**
FLICKR25K	SPQ [17]	71.8	74.0	74.5
	SSCQ (ours)	**73.8**	**75.9**	**76.7**

Table 1: Comparison with SOTA deep fully unsupervised methods on CIFAR-10, NUS-WIDE and FLICKR25K in terms of mAP (%). Some results are cited from [17, 34].

4.2 Comparison with the State of the Art

We denote two variants of our method as SSCQ (fully unsupervised) and SSCQ-p (pre-trained unsupervised), and compare with state-of-the-art unsupervised image retrieval methods, including **(i)** shallow methods with input features extracted from an ImageNet pre-trained VGG16 model [33], such as SpectralH [36] and ITQ [10]; **(ii)** deep pre-trained unsupervised methods which use an ImageNet pre-trained VGG16 model [33] as the backbone and optimize certain layers to generate codes in an unsupervised learning manner, such as Bi-half [24] and MeCoQ [34]; **(iii)** deep fully unsupervised methods which train a model from scratch and jointly optimize visual features and codes in an unsupervised learning manner, such as SPQ [17]. Due to the space limit, we highlight the comparison to deep fully unsupervised methods in Table 1 and refer to Supp. Mat. for the full results.

On CIFAR-10, SSCQ achieves the best performance on all bits. It improves the second-best SPQ by a margin between 1.5% to 2.0%. On the large-scale multi-label NUS-WIDE dataset, SSCQ outperforms SPQ by 0.5% to 3.0%. On the relatively small-scale FLICKR25K dataset, as shown in our SSCQ improves SPQ approximately by 2% on all bits.

In Fig. 3, we report PR curves and P@1000 curves. It can be observed that our SSCQ (blue curve) consistently outperforms SPQ (green curve) under the fully unsupervised setting, while our SSCQ-p (orange curve) performs competitively against the state-of-the-art pre-trained methods. This further demonstrate that our approach is capable of learning effective embeddings and codes for image retrieval at different required recall rates and numbers of top returned samples.

4.3 Coupling part loss with global losses

Our part neighbour loss $\mathcal{L}_{pn}$ is not limited to a specific choice of global loss and we hypothesize it is compatible with any instance-level global loss. Now, we conduct experiments to validate the assumption. We consider four types of global loss in this experiment, and they are quantized loss ($\mathcal{L}_{icz}$), embedding loss ($\mathcal{L}_{icf}$), combined loss ($\mathcal{L}_{icz}+\mathcal{L}_{icf}$) and cross quantized loss in SPQ [17]. Models are trained up to 800 epochs on CIFAR-10, with $\lambda_{cd}=\lambda_{cc}=0$.

In Table 2, we make two observations: **(i)** For all types of global loss, the retrieval performance could be improved when our proposed $\mathcal{L}_{pn}$ is added. Note that, using $\mathcal{L}_{icf}$ alone gives

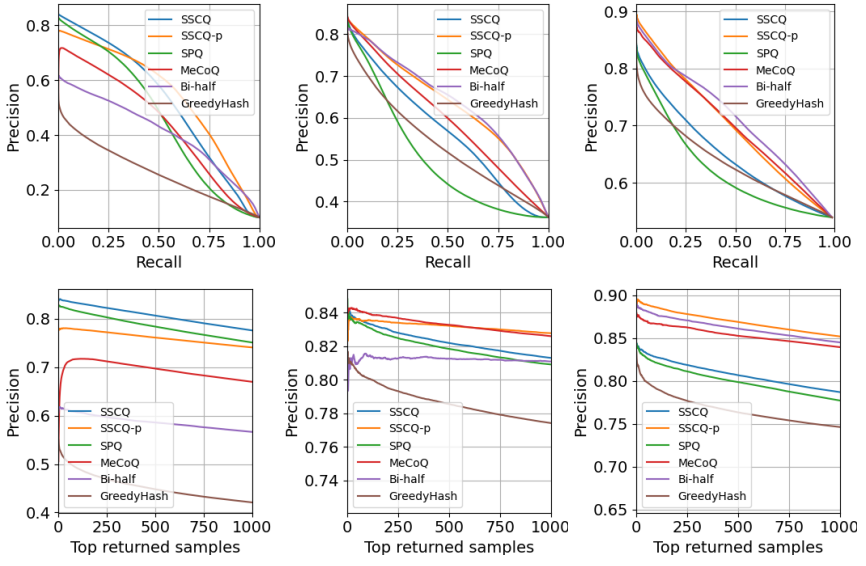


Figure 3: PR curves (*Top*) and P@1000 curves (*Bottom*) on CIFAR-10, NUS-WIDE and FLICKR25K (32 bits).

Global Loss	$\mathcal{L}_{pn}$	mAP(%)↑	SimPos↑	SimNeg↓	Margin↑
$\mathcal{L}_{icz}$	-	74.48	0.68	0.09	0.59
	✓	77.25	0.72	0.10	0.62
$\mathcal{L}_{icf}$	-	10.59	0.29	-0.01	0.30
	✓	76.11	0.29	-0.03	0.32
$\mathcal{L}_{icz}+\mathcal{L}_{icf}$	-	76.28	0.30	-0.03	0.33
	✓	78.64	0.30	-0.03	0.33
SPQ[□]	-	74.73	0.32	-0.03	0.35
	✓	74.96	0.32	-0.04	0.36

Table 2: Different global contrastive losses benefit from the proposed part neighbour semantic loss $\mathcal{L}_{pn}$ on CIFAR-10 (16 bits). We also show the average similarity for positive and negative pairs of the validation set (λ_{pn} is set to 0.1 when $\mathcal{L}_{pn}$ is enabled).

very low mAP, while the mAP could be boosted from 10.59% to 76.11% when the part loss is used. **(ii)** We compute the similarity score between each query and its positive samples and negative samples for the validation set. The margin could be increased after applying $\mathcal{L}_{pn}$, and this verifies that the design of $\mathcal{L}_{pn}$ is successful to learn more discriminative quantized representations.

4.4 Ablation Study

In Table 3, we present component effectiveness evaluation of the proposed SSCQ on CIFAR-10. We make several observations below: **(i)** the part consistent quantization ($\mathcal{L}_{icz}$+$\mathcal{L}_{pn}$+$\mathcal{L}_{cd}$) improves the baseline (top row) by a large margin. **(ii)** The global consistent quantization ($\mathcal{L}_{icz}$+$\mathcal{L}_{icf}$+$\mathcal{L}_{cc}$) also significantly improves the baseline. **(iii)** SSCQ with the unified learning objective (bottom row) of hierarchical consistent quantization yields the best performance, which improves the baseline by 4.0% on average across all bits.

Loss Term					mAP (%)		
$\mathcal{L}_{icz}$	$\mathcal{L}_{pn}$	$\mathcal{L}_{cd}$	$\mathcal{L}_{icf}$	$\mathcal{L}_{cc}$	16 bits	32 bits	64 bits
✓	-	-	-	-	74.2	77.6	78.5
✓	✓	-	-	-	77.3	79.2	80.8
✓	✓	✓	-	-	77.9	80.6	81.9
✓	-	-	✓	-	76.5	80.0	80.8
✓	-	-	✓	✓	76.8	80.2	81.4
✓	✓	✓	✓	✓	**78.3**	**81.3**	**82.9**

Table 3: Component effectiveness evaluation on CIFAR-10

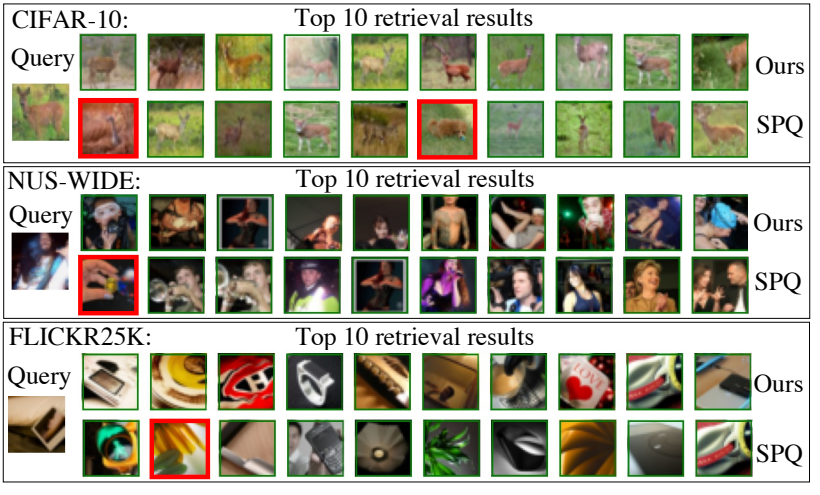


Figure 4: Retrieval results of our approach and SPQ on CIFAR-10, NUS-WIDE and FLICKR25K (32 bits). False retrieval results are denoted in red bounding boxes.

4.5 Qualitative Visualizations

We also visualize some retrieval results of our SSCQ and SPQ [17] in Fig. 4. We can see that both SSCQ and SPQ can retrieve visually similar images from the database, but SSCQ is capable of exploring more discriminative information and results in more relevant retrieval results with higher accuracy.

5 Conclusions

We propose a hierarchical Self-Supervised Consistent Quantization (SSCQ) approach to deep fully unsupervised image retrieval. To exploit self-supervised learning for unsupervised image retrieval at different levels, we devise *part consistent quantization* using part neighbor semantic consistency learning and *global consistent quantization* with consistency regularization. Extensive experiments demonstrate the superior performance of our approach over the state-of-the-art methods. In future work, we aim to explore multi-level hierarchical self-supervision information to facilitate unsupervised cross-modal retrieval, e.g. text to image. It is also promising to incorporate large pre-trained language-aligned visual encoders such as CLIP [29] into our method for both uni-modal or multi-modal applications.

References

[1] Artem Babenko and Victor Lempitsky. Additive quantization for extreme vector compression. In *Proceedings of the IEEE Conference on CVPR*, pages 931–938, 2014.

[2] Yue Cao, Mingsheng Long, Jianmin Wang, Han Zhu, and Qingfu Wen. Deep quantization network for efficient image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3457–3463, 2016.

[3] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Philip S Yu. Hashnet: Deep learning to hash by continuation. In *Proceedings of the IEEE ICCV*, pages 5608–5617, 2017.

[4] Moses S Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 380–388, 2002.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.

[6] 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 *Proceedings of the ACM international conference on image and video retrieval*, pages 1–9, 2009.

[7] Bo Dai, Ruiqi Guo, Sanjiv Kumar, Niao He, and Le Song. Stochastic generative hashing. In *International Conference on Machine Learning*, pages 913–922. PMLR, 2017.

[8] Kamran Ghasedi Dizaji, Feng Zheng, Najmeh Sadoughi, Yanhua Yang, Cheng Deng, and Heng Huang. Unsupervised deep generative adversarial hashing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3664–3673, 2018.

[9] Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun. Optimized product quantization for approximate nearest neighbor search. In *Proceedings of the IEEE CVPR*, pages 2946–2953, 2013.

[10] Yunchao Gong, Svetlana Lazebnik, Albert Gordo, and Florent Perronnin. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12): 2916–2929, 2012.

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[12] Jae-Pil Heo, Youngwoon Lee, Junfeng He, Shih-Fu Chang, and Sung-Eui Yoon. Spherical hashing. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2957–2964. IEEE, 2012.

[13] Jiabo Huang, Qi Dong, Shaogang Gong, and Xiatian Zhu. Unsupervised deep learning by neighbourhood discovery. In *International Conference on Machine Learning*, pages 2849–2858. PMLR, 2019.

[14] Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international Conference on Multimedia Information Retrieval*, pages 39–43, 2008.

[15] Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2785–2795, 2022.

[16] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613, 1998.

[17] Young Kyun Jang and Nam Ik Cho. Self-supervised product quantization for deep unsupervised image retrieval. In *Proceedings of the IEEE ICCV*, pages 12085–12094, 2021.

[18] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33 (1):117–128, 2010.

[19] Benjamin Klein and Lior Wolf. End-to-end supervised product quantization for image search and retrieval. In *Proceedings of the IEEE CVPR*, pages 5041–5050, 2019.

[20] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018.

[21] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, 2009.

[22] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*, 2021.

[23] Qi Li, Zhenan Sun, Ran He, and Tieniu Tan. Deep supervised discrete hashing. *Advances in Neural Information Processing Systems*, 30, 2017.

[24] Yunqiang Li and Jan van Gemert. Deep unsupervised image hashing by maximizing bit entropy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2002–2010, 2021.

[25] Wei Liu, Cun Mu, Sanjiv Kumar, and Shih-Fu Chang. Discrete graph hashing. *Advances in Neural Information Processing Systems*, 27, 2014.

[26] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[27] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

[28] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3): 145–175, 2001.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[30] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015.

[31] Ruslan Salakhutdinov and Geoffrey Hinton. Semantic hashing. *International Journal of Approximate Reasoning*, 50(7):969–978, 2009.

[32] Yuming Shen, Jie Qin, Jiaxin Chen, Mengyang Yu, Li Liu, Fan Zhu, Fumin Shen, and Ling Shao. Auto-encoding twin-bottleneck hashing. In *Proceedings of the IEEE CVPR*, pages 2818–2827, 2020.

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[34] Jinpeng Wang, Ziyun Zeng, Bin Chen, Tao Dai, and Shu-Tao Xia. Contrastive quantization with code memory for unsupervised image retrieval. In *Proceedings of the AAAI*, 2022.

[35] Chen Wei, Huiyu Wang, Wei Shen, and Alan Yuille. Co2: Consistent contrast for unsupervised visual representation learning. In *International Conference on Learning Representations*, 2021.

[36] Yair Weiss, Antonio Torralba, and Rob Fergus. Spectral hashing. *Advances in Neural Information Processing Systems*, 21, 2008.

[37] Guile Wu, Xiatian Zhu, and Shaogang Gong. Tracklet self-supervised learning for unsupervised person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12362–12369, 2020.

[38] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[39] Erkun Yang, Tongliang Liu, Cheng Deng, Wei Liu, and Dacheng Tao. Distillhash: Unsupervised deep hashing by distilling data pairs. In *Proceedings of the IEEE CVPR*, pages 2946–2955, 2019.

[40] Qize Yang, Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Patch-based discriminative feature learning for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3633–3642, 2019.

[41] Tan Yu, Jingjing Meng, Chen Fang, Hailin Jin, and Junsong Yuan. Product quantization network for fast visual search. *International Journal of Computer Vision*, 128(8):2325–2343, 2020.

[42] Li Yuan, Tao Wang, Xiaopeng Zhang, Francis EH Tay, Zequn Jie, Wei Liu, and Jiashi Feng. Central similarity quantization for efficient image and video retrieval. In *Proceedings of the IEEE CVPR*, pages 3083–3092, 2020.

[43] Maciej Zieba, Piotr Semberecki, Tarek El-Gaaly, and Tomasz Trzcinski. Bingan: Learning compact binary descriptors with a regularized gan. *Advances in Neural Information Processing Systems*, 31, 2018.