Cross-Quantized Hyperbolic Representations for Enhancing Cartoon Image Retrieval

Thanh-Hung Nguyen[1,2] and Thi-Ngoc-Hanh Le[1,2⋆]

[1] International University, VNU-HCM, Ho Chi Minh City, Vietnam
[2] Vietnam National University Ho Chi Minh City, Viet Nam
hungnguyenarbeit@gmail.com, ltnhanh@hcmiu.edu.vn

Abstract. Approximate Nearest Neighbor search via Product Quantization has become a popular approach for large-scale unsupervised image retrieval. However, existing systems primarily focus on real-world domains and lack evaluation in animation and cartoon contexts, where hierarchical semantics structure plays a crucial role in creative workflows. In animation production, artists often require flexible retrieval tools that can suggest visually coherent yet stylistically diverse content based on a given frame. To address this gap, we propose Cross-Quantized Hyperbolic (CQH) retrieval approach, which can capture hierarchical semantic similarity within the context of hyperbolic geometry. Besides, we contribute a so-called Cartoon18K dataset, the first large-scale dataset of cartoon images with multi-label hierarchical annotations. Extensive experiments on Flickr25K, NUS-WIDE, and CIFAR-10 demonstrate that CQH outperforms state-of-the-art unsupervised baselines. This accomplishment underscores the efficacy of our approach and contributes to the advancement of cartoon retrieval amidst the rapid proliferation of creative multimedia content. Our code and dataset are available at: https://github.com/Greekatz/CQH.

Keywords: CQH · Cross Quantized Hyperbolic · Cartoon Retrieval · Hyperbolic Product Quantization

1 Introduction

Image retrieval [25, 9] is the task of finding images visually or semantically similar to a query image from a large database. It is a fundamental technology for many industrial applications such as visual search engines [15], digital asset management systems [9], or personalized image recommendation systems [2]. Applications of image retrieval in animation and entertainment include reusing cartoon characters, referring to styles, and indexing assets of cartoon video archives [16, 45]. Cartoon retrieval is challenging as cartoon images have abstract properties, exaggerated and varied geometric shapes, and are different from natural images. To handle large-scale databases, unsupervised retrieval methods perform an Approximate Nearest Neighbor (ANN) search without relying on labeled data. ANN

⋆ Corresponding author: ltnhanh@hcmiu.edu.vn

are primarily categorized into Binary Hashing (BH) [35, 17] and Product Quantization (PQ) [21, 46]. BH techniques encode data into compact binary codes, enabling fast similarity searches via Hamming distance, but they often suffer from information loss due to integer-only distance values. Conversely, PQ divides the high-dimensional space and estimates distances utilizing pre-calculated real-valued inter-codeword distances, hence demonstrating greater similarity, and has been applied to video search [47].

Most existing unsupervised image retrieval systems [27, 20, 42, 33] have achieved success on large-scale datasets with a wide range of real-world concepts. However, they lack evaluation on animation styles and cartoon scenes. In animation production, professional artists must ensure that the generated content is plausible in both appearance and motion direction. Although existing commercial applications such as Toon Boom [41] and Adobe Animate [1] can handle character animation, they offer limited diversity in animation styles and cartoon scenes [24]. Therefore, it is necessary to develop a cartoon retrieval system that enables fast and interactive animation creation based on user preferences, allowing artists to focus more on creative work.

To the end, our goal in this research is to develop a retrieval system capable of reflecting multi-level similarity based on a given query. The need for modeling multi-level semantic relationships between images has been directly addressed by HiHPQ [33]. Their work learns quantized representations by incorporating hierarchical semantic similarity within hyperbolic geometry. Nonetheless, despite its alignment with our motivation, we identify two fundamental challenges in its formulation that limit both training stability and retrieval performance, preventing it from fully realizing the potential of hierarchical retrieval in cartoon preference scenarios.

First, the hyperbolic product quantization module is used during initial training. At this stage, the quantized features are not yet semantically aligned with the continuous features. When this misalignment interacts with learning curvature optimization, the quantizer may form codebooks in suboptimal regions of the manifold, leading to distorted hierarchical structures and inconsistent codeword assignments.

Second, contrastive learning (CL) suffers from a trade-off between effectiveness and efficiency. A large number of negative pairs is necessary to maintain strong discrimination among samples, but a single GPU memory restricts the feasible batch size. Consequently, training with CL either leads to reduced representation quality with small batches or increased computational cost when scaling across multiple GPUs.

Therefore, we propose **Cross-Quantized Hyperbolic (CQH)**, a framework that learns both quantized and continuous representations through a hyperbolic product quantizer module. Our primary objective is to mitigate representation distortion by jointly learning both representations within a shared hyperbolic space. The continuous representations provide a stable, high-fidelity target to guide the learning of the quantizer, especially in the early epochs. Furthermore, cross-quantized contrastive learning introduces additional samples from the con-

tinuous space, thereby alleviating the sampling bias inherent in in-batch negative sampling.
We evaluate our proposed framework on standard benchmarks: Flickr25K [19], NUS-WIDE [7], CIFAR-10 [23], our newly created dataset, dubbed as Cartoon18K. To the best of our knowledge, Cartoon18K is the first public cartoon domain dataset with multi-label annotations designed for unsupervised image retrieval. We plan to maintain and grow this dataset to support the research community in the near future.
The contributions of our work are summarized as follows:

- We identify the critical problem of instability in the hyperbolic product quantizer. To resolve this problem, we propose a novel framework that stabilizes hyperbolic quantization and mitigates sampling bias through hyperbolic cross-quantized contrastive learning.
- We develop the **Cartoon18K**, a novel dataset with multi-label hierarchical annotations to address the evaluation gap in complex, structured domains.
- Extensive experiments on several benchmarks demonstrate that our proposed CQH method outperforms state-of-the-art baselines.

2 Related Work

2.1 Binary Hashing

One of the most common hashing algorithms is Locality Sensitive Hashing (LSH) [5], it utilizes random linear projection to measure the similarity between different bit codes. Semantic Hashing (SemH) [35] is another popular binary hashing method that learn a graphical model to capture similar documents in a time that is independent of the document length. Later, deep learning based binary hashing methods have been proposed to learn binary codes from raw data. Cao et al. [4] proposed a hashNet that learns binary codes from raw data by training a deep neural network to map the input data to a binary code. Li et al. [26] proposed a deep supervised discrete hashing algorithm that combines both the pairwise label information and the similarity between the binary codes to learn a better binary code. Although these supervised methods work well, they rely on data annotations done by humans, which are expensive and time-consuming to obtain. To resolve this problem, researchers have resorted to deep unsupervised hashing for image retrieval. Existing unsupervised deep hashing methods can be categorized into generative [8, 11, 36] or discriminative approaches [29, 18, 44]. Generative methods can either employ autoencoders [22], which expect the binary codes to reconstruct the original input data, or utilize generative adversarial networks (GANs) [14] to reconstruct likelihood representation through the discriminator. On the other hand, discriminative methods focus on preserving the similarity among continuous feature representations in the learned Hamming space. Typically, they heavily depend on high-quality pretrained features. If a pretrained CNN cannot generalize well to new domains, the adopted CNN may extract unsatisfactory features, fundamentally harming the generated hash codes.

2.2 Product Quantization

Recently, there have been many end-to-end differentiable unsupervised product quantization frameworks with impressive performance introduced. Morozov et al. [30] revealed an unsupervised product quantization framework that exploited reconstruction loss to learn product representations without labels. Jang et al. [20] proposed a self-supervised product quantization (SPQ) framework via cross-quantized contrastive learning to maximize the cross similarity between the same images from both deep descriptors and product quantized descriptors. In a related direction, Wang et al. [42] introduced a codeword diversity regularizer to prevent model degeneration and utilized a code memory to enhance contrastive learning with lower feature drift. Most recently, Qiu et al. [33] proposed a hyperbolic product quantization (HiHPQ) that hierarchically produces multilevel codebooks. Specifically, they proposed a semantics learning module from prototype learning to capture a better tree-like structure in a hyperbolic space. However, SPQ only optimized the problem of deep product quantization in Euclidean space, neglecting the underlying hierarchical semantic structure of data, which is prevalent in real-world applications. HiHPQ, on the other hand, suffers from instability during the initial training phase due to unstable quantized features and sampling bias in contrastive learning.

3 Methodology

3.1 Preliminary

Lorentz Model. Lorentz model [31], also known as the hyperboloid model, it depicts the d-dimensional hyperbolic space (*i.e.*, Lorentzian manifold) as a submanifold of $\mathbb{R}^{d+1}$. The d-dimensional Lorentz model with constant negative curvature $-\theta$ is defined as the following set of vectors:

$$\mathbb{H}_\theta^d = \left\{ x \in \mathbb{R}^{d+1} \;\middle|\; \langle x, x \rangle_{\mathcal{L}} = -\frac{1}{\theta},\ \theta > 0 \right\}. \tag{1}$$

Lorentzian Distance. The Lorentzian distance between x, y $\in \mathbb{H}_\theta^d$ is defined as:

$$d_{\mathbb{H}_\theta^d}(x, y) = \frac{1}{\sqrt{\theta}} \operatorname{arcosh}(-\theta \langle x, y \rangle_{\mathcal{L}}), \tag{2}$$

which depicts the length of the shortest path (i.e., geodesic) between x and y on the hyperboloid. With $\theta \to 0$, the Lorentzian distance in 1 converges to the Euclidean distance.

Tangent Space. For any vector $p \in \mathbb{H}_\theta^d$, the tangent space $T_p\mathbb{H}_\theta^d$ at point p is defined as the first-order approximation of the manifold around p:

$$T_p\mathbb{H}_\theta^d = \left\{ v \in \mathbb{R}^{d+1} \;\middle|\; \langle p, v \rangle_{\mathcal{L}} = 0 \right\} \tag{3}$$

For any ambient Euclidean vector $u \in \mathbb{R}^{d+1}$, its orthogonal projection onto the tangent space $T_p\mathbb{H}^d_\theta$ is given by:

$$\text{proj}_p(u) = u + \theta p \langle p, u \rangle_{\mathcal{L}}. \tag{4}$$

Exponential Map. The exponential map $\exp_p : T_p\mathbb{H}^d_\theta \to \mathbb{H}^d_\theta$ maps a tangent vector $v \in T_p\mathbb{H}^d_\theta$ to the manifold $\mathbb{H}^d_\theta$ along the geodesic starting at p with initial velocity v:

$$\exp_p(v) = \cosh(\sqrt{\theta}\|v\|_{\mathcal{L}})p + \frac{\sinh(\sqrt{\theta}\|v\|_{\mathcal{L}})}{\sqrt{\theta}\|v\|_{\mathcal{L}}}v, \tag{5}$$

where $\|v\|_{\mathcal{L}} = \sqrt{\langle v, v \rangle_{\mathcal{L}}}$ is the Lorentzian norm of v.

3.2 System Overview

In our current work, we employ hierarchical hyperbolic product quantization [33] as the baseline. Our proposed Cross-Quantized Hyperbolic (CQH) model aims to enhance representation learning and retrieval performance for cartoon image by integrating geometry, quantized embeddings, and contrastive learning. The overall pipeline of CQH is illustrated in Fig.1.
Given a set of cartoon images x, the framework first extracts deep features f_x via a shared feature extraction backbone. In our experiment, we use the pre-trained VGG-16 [38] model, other backbones (such as VGG-19, ResNet, *etc.*,) can result in equivalent effect. The shared encoder supports the semantically similar cartoon samples are embedded in comparable Euclidean feature spaces. The extracted Euclidean feature f_x are then mapped onto hyperbolic manifold $\mathbb{H}^d_\theta$ via the exponential mapping (ExpMap) to obtain hyperbolic representations, $h_x = \exp(f_x)$. This mapping helps the model to encode hierarchical and tree-like semantic relationships in visual cartoon representations (such as, character relationships, shapes, or style). The h_x is then processed through by the hyperbolic product quantizer to produce quantized embeddings $\hat{h}_x$. This step captures local geometric structure while preserving the global curvature properties of the manifold. Finally, both the continuous embedding h_x and the quantized embedding $\hat{h}_x$ are fused via our proposed hyperbolic cross-quantized contrastive learning. In this phase, the anchor representation is constructed by combining h_x and $\hat{h}_x$. The anchor contrasts with both instance and prototype representations (z_i, β_i) in the hyperbolic space using Lorentzian distance metrics. To train the feature extraction and the hyperbolic quantization head in an end-to-end manner, soft quantization mechanism [46] is employed for training.

3.3 Proposed Method

Hyperbolic Cross-Quantized Contrastive Learning. Inspired by recent success of [20], we introduce a hyperbolic cross-quantized contrastive learning

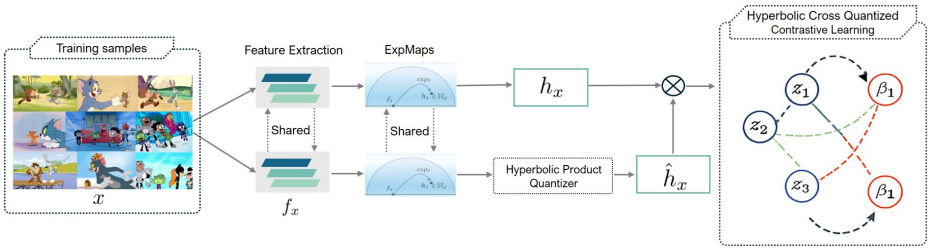


Fig. 1. Visualization of our CQH framework.

scheme to jointly optimize both the encoder network and quantization module through cross-space interactions. Given a mini-batch of training images $\mathcal{B}$ with batch size N_B, we apply random augmentations to each image twice, resulting in $2N_B$ views $\{(x_i^{(1)}, x_i^{(2)})\}_{i=1}^{N_B}$. After encoding and quantization, we obtain continuous hyperbolic representations $\{h_{x_i}^{(1)}, h_{x_i}^{(2)}\}_{i=1}^{N_B}$ and quantized hyperbolic representations $\{\hat{h}_{x_i}^{(1)}, \hat{h}_{x_i}^{(2)}\}_{i=1}^{N_B}$ composed of sub-vectors quantized across multiple hyperbolic codebooks.

Let $X_a, X_b \in \mathbb{H}^{N_B \times M \times D}$ be the continuous hyperbolic embeddings from two augmented views, and $Z_a, Z_b \in \mathbb{H}^{N_B \times M \times D}$ be the corresponding quantized embeddings, where M is the number of product components and D is the dimension of each subspace. We define the cross-quantized embeddings as:

$$X_a Z_b = [X_a; Z_b] \in \mathbb{H}^{2N_B \times M \times D} \tag{6}$$

$$X_b Z_a = [X_b; Z_a] \in \mathbb{H}^{2N_B \times M \times D} \tag{7}$$

The pairwise similarity is defined as the negative sum of Lorentzian distances across all M components:

$$\mathrm{Sim}_{i,j} = -\sum_{m=1}^{M} d_L(x_i^{(m)}, x_j^{(m)}; \theta_m) \tag{8}$$

where $d_L(\cdot, \cdot; \theta)$ is the Lorentzian distance under curvature θ_m for the m-th subspace.

The Hyperbolic Cross-Quantized Contrastive (HCQC) loss is then:

$$\mathcal{L}_{\mathrm{HCQC}} = \frac{1}{2N_B} \sum_{i=1}^{2N_B} \left[-\log \frac{\exp(\mathrm{Sim}_{i,i+N_B}^{ab}/\tau)}{\sum_{j=1}^{2N_B} \exp(\mathrm{Sim}_{i,j}^{ab}/\tau)} - \log \frac{\exp(\mathrm{Sim}_{i,i+N_B}^{ba}/\tau)}{\sum_{j=1}^{2N_B} \exp(\mathrm{Sim}_{i,j}^{ba}/\tau)} \right] \tag{9}$$

where τ is a temperature parameter, and the superscripts ab and ba indicate the cross-quantized similarity computation between continuous and quantized representations.

Hierarchical Semantic Learning Module. To capture hierarchical semantic structure, we adopt the hierarchical semantic learning module from HiHPQ [33].

This module employs bottom-up hierarchical clustering in the Euclidean tangent space $\mathcal{T}_{\mathbf{p}}\mathbb{S}$ to extract multi-level semantic structures, then maps prototypes to the hyperbolic product manifold via exponential mapping for interaction with quantized representations.
The baseline combines prototype-wise and instance-wise contrastive learning:

$$\mathcal{L}_{\text{proto}} = -\sum_{i=1}^{N_B} \frac{1}{L} \sum_{l=1}^{L} \log \frac{\mathcal{S}(\hat{\mathbf{h}}_{x_i}, \tilde{\mathbf{e}}_j^l(i))}{\sum_{n=1}^{N_l} \mathcal{S}(\hat{\mathbf{h}}_{x_i}, \tilde{\mathbf{e}}_n^l)} \tag{10}$$

$$\mathcal{L}_{\text{neighbor}} = -\sum_{i=1}^{N_B} \frac{1}{L} \sum_{l=1}^{L} \log \frac{\mathcal{S}(\hat{\mathbf{h}}_{x_i}, \hat{\mathbf{h}}_{x_{il}})}{\sum_{t \in \mathcal{B} \setminus x_i} \mathcal{S}(\hat{\mathbf{h}}_{x_i}, \hat{\mathbf{h}}_t)} \tag{11}$$

where $\mathcal{S}(\cdot, \cdot)$ is the negative hyperbolic distance, $\tilde{\mathbf{e}}_j^l(i)$ is the prototype for image x_i at hierarchy level l, and x_{il} is a randomly selected image from the same prototype group.
The hierarchical semantic loss is:

$$\mathcal{L}_{\text{HS}} = \lambda_2 \mathcal{L}_{\text{proto}} + \lambda_3 \mathcal{L}_{\text{neighbor}}, \tag{12}$$

where λ_2 and λ_3 control the relative importance of the two components.
Overall, the training objective is defined as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{HCQC}} + \mathcal{L}_{\text{HS}}, \tag{13}$$

The complete training procedure for our CQH framework is outlined in Algorithm 1. The algorithm jointly optimizes our proposed hyperbolic cross-quantized contrastive learning with the baseline hierarchical semantic learning modules.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our framework on four benchmarks: Flickr25K [19], CIFAR-10 [23], NUS-WIDE [7], and our newly collected Cartoon18K dataset. Flickr25K consists of 25,000 images across 24 semantic categories. Following the setting in [33], we randomly sample 2,000 images as queries and use 5,000 images from the remaining set for training. CIFAR-10 contains 60,000 images from 10 classes. We adopt two evaluation protocols. In Protocol I, 10,000 images (1,000 per class) are used as queries and the other 50,000 images serve as the training set. In Protocol II, 1,000 images per class are used as queries and 500 images per class are used for training. In both settings, the retrieval database includes all images excluding the query set. NUS-WIDE includes about 270k web images annotated with 81 categories. Following common practice, we evaluate on the 21 most frequent categories. From these, 100 images per category are used as queries, and the remaining images form the retrieval database. Additionally, we sample 500 images per category for training, resulting in 10,500 training examples.

Algorithm 1 CQH Training Algorithm

Require: Unlabeled training data $\mathcal{X}$, model $\mathcal{M}$, warmup epochs E_w, clustering interval Δ
Ensure: Trained model for image retrieval

for epoch $e = 1$ to $E_{\max}$ **do**
  if $e > E_w$ **and** $(e - E_w) \mod \Delta = 0$ **then**
    /* Hierarchical clustering phase */
    Extract features $\{\mathbf{h}_i\}$ from all training samples
    Perform hierarchical clustering to obtain cluster assignments $\mathcal{C}^{(l)}$ at each level l
    Store image-to-cluster mappings: im2cluster$^{(l)}$ for all levels
  end if
  for sampled mini-batch $\{x_i\}_{i=1}^{N}$ **do**
    Generate two augmented views $(x_i^{(1)}, x_i^{(2)})$ for each x_i
    Extract continuous hyperbolic representation $\mathbf{h}$ of all views
    Extract quantized hyperbolic representation $\hat{\mathbf{h}}$ of all views
    /* Our proposed HCQC learning */
    if im2cluster exists **then**
      Compute $\mathcal{L}_{\text{HCQC}}$ with neighbor-aware sampling using im2cluster (Eq. 9)
    else
      Compute $\mathcal{L}_{\text{HCQC}}$ without hierarchical structure
    end if
    /* Hierarchical semantic learning */
    if im2cluster exists **then**
      Compute $\mathcal{L}_{\text{HS}}$ using cluster centroids and assignments (Eq. 12)
    end if
    /* Unified learning objective */
    Optimize model with $\mathcal{L} = \mathcal{L}_{\text{HCQC}} + \lambda \mathcal{L}_{\text{HS}}$
  end for
end for

Cartoon18K Dataset. We introduce a novel Cartoon18K dataset to address the underexplored of unsupervised retrieval domain of cartoon media. The Cartoon18K dataset was gathered by downloading 10 cartoon YouTube videos that encompass three timeless series with varied artistic styles: *Teen Titans Go!*, *Tom & Jerry*, and *Looney Tunes*. We took 1,800 keyframes per video from the videos to get a pool of 18,000 images altogether. Each image has been manually annotated with multi-labels spanning over 15 categories including characters, actions, and scenes. Following the same sampling protocol as Flickr25K, 2,000 images are randomly selected as testing queries, while 5,000 images are randomly chosen as the training set from the remaining images. The retrieval database includes all non-query images.

4.2 Implementation Details

Following [33], the encoder is constituted with the pre-trained VGG-16 network followed by a linear projector. We set the initial learning rate to 1×10^{-3} and

Table 1. MAP (%) comparison between unsupervised retrieval methods."HPQ" denotes Hyperbolic Product Quantization; "PQ" denotes Product Quantization; "BH" denotes Binary Hashing.

Method	Type	Flickr25K			CIFAR-10 (I)			CIFAR-10 (II)			NUS-WIDE		
		16bits	32bits	64bits	16bits	32bits	64bits	16bits	32bits	64bits	16bits	32bits	64bits
LSH+VGG	BH	56.11	57.08	59.26	14.38	15.86	18.09	12.55	13.76	15.07	38.52	41.43	43.89
SH+VGG	BH	59.77	61.36	64.08	27.09	29.44	32.65	27.20	28.50	30.00	51.70	51.10	51.00
SpH+VGG	BH	61.32	62.47	64.49	26.90	31.75	35.25	25.40	29.10	33.30	49.50	55.80	58.20
ITQ+VGG	BH	63.30	65.92	68.86	34.41	35.41	38.82	30.50	32.50	34.90	62.70	64.50	66.40
DeepBit	BH	62.04	66.54	68.34	19.43	24.86	27.73	20.60	28.23	31.30	39.20	40.30	42.90
SGH	BH	72.10	72.84	72.83	34.51	37.04	38.93	43.50	43.70	43.30	59.30	59.00	60.70
HashGAN	BH	72.11	73.25	75.46	44.70	46.30	48.10	42.81	47.54	47.29	68.44	70.56	71.71
GreedyHash	BH	69.97	70.85	73.03	44.80	47.20	50.10	45.76	48.26	53.34	63.30	69.10	73.10
BinGAN	BH	-	-	-	-	-	-	47.60	51.20	52.00	65.40	70.90	71.30
BGAN	BH	-	-	-	-	-	-	52.50	53.10	56.20	68.40	71.40	73.00
SSDH	BH	75.65	77.10	76.68	36.16	40.17	44.00	33.30	38.29	40.81	58.00	59.30	61.00
DVB	BH	-	-	-	-	-	-	40.30	42.20	44.60	60.40	63.20	66.50
TBH	BH	74.38	76.14	77.87	54.68	58.63	62.47	53.20	57.30	57.80	71.70	72.50	73.50
Bi-half Net	BH	76.07	77.93	78.62	56.10	57.60	59.50	49.97	52.04	55.35	76.86	78.31	79.94
CIBHash	BH	77.21	78.43	79.59	59.39	63.67	65.16	59.00	62.20	64.10	79.00	80.70	81.50
PQ+VGG	PQ	62.75	66.63	69.40	27.14	33.30	37.67	28.16	30.24	30.61	65.39	67.41	68.56
OPQ+VGG	PQ	63.27	68.01	69.86	27.29	35.17	38.48	32.17	33.50	34.46	65.74	68.38	69.12
DeepQuan	PQ	-	-	-	39.95	41.25	43.26	-	-	-	-	-	-
SPQ	PQ	77.35	78.74	79.98	63.17	66.88	68.02	56.56	61.45	63.30	78.51	80.41	81.70
HiHPQ	HPQ	80.74	80.45	80.76	**72.66**	72.43	70.45	63.33	65.17	66.45	**80.18**	81.06	81.92
Ours	HPQ	**81.03**	**81.06**	**82.33**	71.19	**73.54**	**72.03**	**63.54**	**66.92**	**68.35**	79.86	**81.51**	**82.02**

Table 2. MAP (%) results of our framework on **Cartoon18K**.

Method	Cartoon18K		
	16 bits	32 bits	64 bits
Ours	79.02	77.35	77.42

decreased it in a cosine decay manner until it reached 1e-5. We apply data augmentation techniques, including random cropping, horizontal flipping, image graying, random color distortions, and Gaussian blur. During training, the batch size is set to 128 and the total number of epochs is 50. The learning curvature parameter $-\theta_i$ for each Lorentzian manifold is initialized to 1.0. Our model is implemented in PyTorch [32] and trained on a single P100 GPU. We adopt Riemannian SGD [3] as the optimizer. The number of codewords K in each codebook C^m is fixed to 256, and the dimension of each codebook is set to 16. By setting the number of sub-codebooks $M = 2, 4, 8$, we obtain final code lengths of $16, 32, 64$ bits, respectively. For dataset-specific configurations, we set $\tau = 0.2$ for CIFAR-10 (I). For CIFAR-10 (II), we set $\tau = 0.5$ for 32/64-bit and $\tau = 0.2$ for the 16-bit setting. For Flickr25K, NUS-WIDE, and Cartoon18K, we use $\tau = 0.5$. In addition, we set $\lambda_2 = 0.5$ for CIFAR-10 experiments and $\lambda_3 = 0.1$ for all experiments.

4.3 Our Results and Discussions

Evaluation Metrics. We utilize the Mean Average Precision (MAP) at top R (*i.e.*, MAP@R) as the primary evaluation metric for assessing the performance of our proposed method. We adopt the MAP@1000 for CIFAR-10 (I) and CIFAR-10 (II) datasets, while MAP@5000 is employed for the Flickr25K, NUS-WIDE, and Cartoon18K datasets.

Baselines. We consider following unsupervised baselines for comparison: (i) binary hashing methods: LSH [5], SH [43], SpH [17], ITQ [13]; DeepBit [29], SGH [8], HashGAN [10], GreedyHash [40], BinGAN [48], BGAN [39], SSDH [44], DVB [36], TBH [37], Bi-half Net [28], CIBHash [34]; and (ii) product quantization methods: PQ [21], OPQ [12]; DeepQuan [6], SPQ [20]. (iii) hyperbolic product quantization method: HiHPQ[3] [33].

Overall Performance As demonstrated in Table 1, our proposed model surpasses both existing baselines and state-of-the-art product quantization frameworks on public benchmark image datasets. Notably, our method achieves substantial improvements over HiHPQ [33], with increases of 0.61%, 0.61%, and 1.57% in MAP@5000 for Flickr25K at 16, 32, and 64 bits, respectively. However, we observe a slight decline of 1.47% at 16 bits on CIFAR-10 (I) when compared to HiHPQ, which we attribute to the limited representation capacity of 16-bit codes in capturing intricate image semantics. We also evaluate our framework on the Cartoon18K dataset, as detailed in Table 2. Our approach achieves a good MAP@5000 score at 16 bits, while a slight degradation is observed at 32 and 64 bits, indicating that short hash codes can better capture the stylized and domain-specific properties of Cartoon18K.

Qualitative Retrieval Results In Figure 2–3, we visualize our retrieval results returned by HiHPQ [33] and our CQH, which are visually similar to the database samples. Our results show that CQH can better discriminate fine-grained similarities compared to HiHPQ. We also evaluate our model on the **Cartoon18K** in Fig 4–5 to demonstrate that both our model and dataset can provide hierarchical semantics similarity in between frames.

Hyperparameter Sensitivity Analysis We next investigate the sensitivity of the temperature τ on CIFAR-10 (II), as shown in Fig. 6(a). The performance improves steadily and reaches its optimum at $\tau = 0.5$ for the 32-bit and 64-bit settings, while the 16-bit setting performs best at $\tau = 0.2$. For other experiments in this paper, except CIFAR-10, we select τ as 0.5.

[3] As our training setup is different from [33], we reproduce HiHPQ using their codebase.

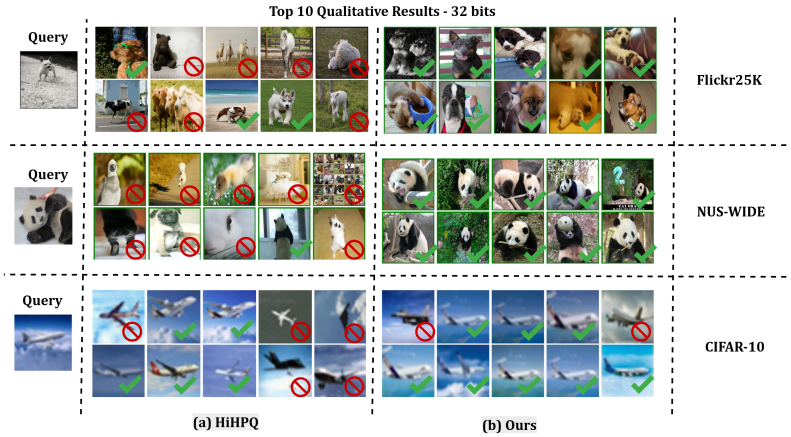


Fig. 2. Qualitative top-10 retrieval results on benchmarks at 32 bits.

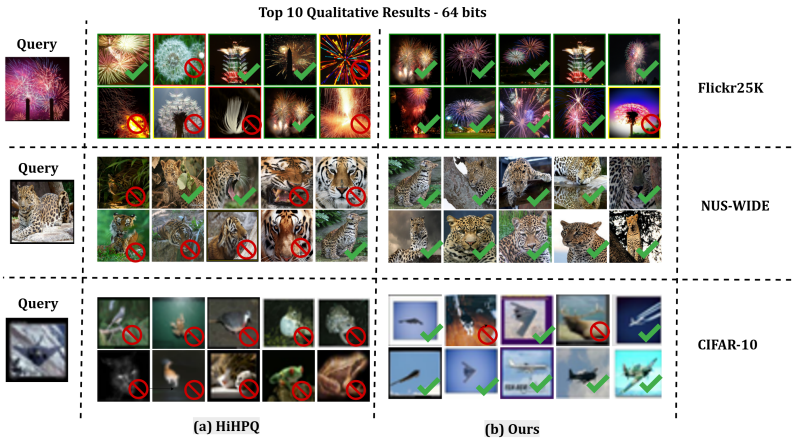
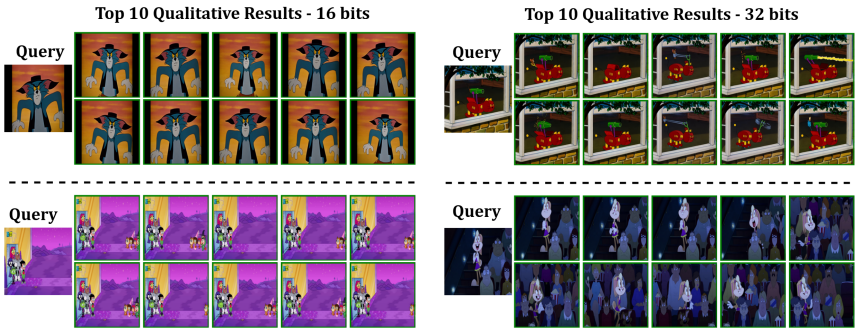


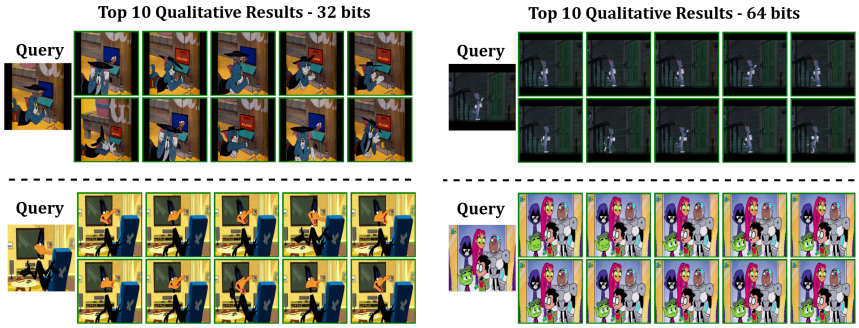
Fig. 3. Qualitative top-10 retrieval results on benchmarks at 64 bits.

Quantization Error Analysis To further evaluate the effectiveness of our proposed framework, we compare the quantization error of CQH with the baseline HiHPQ on the Flickr25K dataset using 32-bit codes, as shown in Fig. 6(b). Here, the quantization error refers to the hyperbolic distance between the continuous representation of an image and its quantized representation. From the results, CQH consistently achieves a lower quantization error than HiHPQ throughout the entire training process and converges more rapidly.



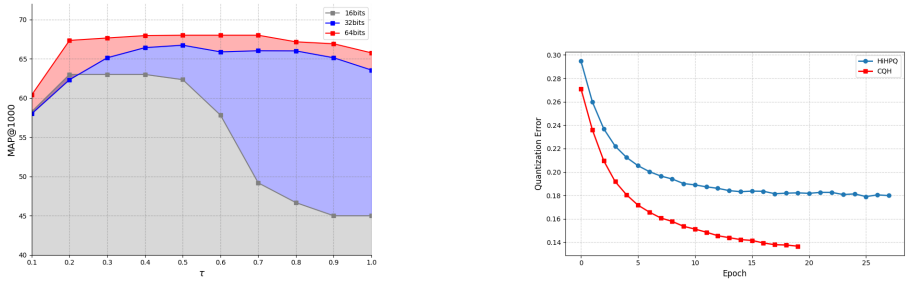
(a) 16 bits (b) 32 bits

Fig. 4. Our top 10 qualitative results of Cartoon18K on different bit lengths. The green border represents as true positives.



(a) 32 bits (b) 64 bits

Fig. 5. Another top 10 qualitative results of Cartoon18K



(a) (b)

Fig. 6. (**a**) Effects of the temperature τ on CIFAR-10 (II). (**b**) Quantization error on Flickr25K (32bits).

5 Conclusion

In this paper, we propose a Cross-Quantized Hyperbolic Representations (CQH) framework to improve cartoon image retrieval. Different from prior work focusing on large-scale benchmarks representing real-world concepts, our CQH targets structured and stylized domains such as animation, where hierarchical semantic structures play a crucial role in creative workflow. Extensive experiments on multiple benchmarks demonstrate that our proposed method outperforms state-of-the-art baselines. In future work, we plan to further develop our dataset for multiple tasks and contribute to research in plausible cartoon animation.

References

1. Adobe Inc.: Adobe animate. Computer software (2024), https://www.adobe.com/products/animate.html, version 24.0.3
2. Ay, B., Aydın, G., Koyun, Z., Demir, M.: A visual similarity recommendation system using generative adversarial networks. In: 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML). pp. 44–48 (2019). https://doi.org/10.1109/Deep-ML.2019.00017
3. Bonnabel, S.: Stochastic gradient descent on riemannian manifolds. IEEE Transactions on Automatic Control **58**(9), 2217–2229 (2013)
4. Cao, Z., Long, M., Wang, J., Yu, P.S.: Hashnet: Deep learning to hash by continuation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
5. Charikar, M.S.: Similarity estimation techniques from rounding algorithms. In: Proceedings of the thiry-fourth annual ACM symposium on Theory of computing. pp. 380–388 (2002)
6. Chen, J., Cheung, W.K., Wang, A.: Learning deep unsupervised binary codes for image retrieval. In: IJCAI. pp. 613–619 (2018)
7. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: 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. pp. 1–9 (2009)
8. Dai, B., Guo, R., Kumar, S., He, N., Song, L.: Stochastic generative hashing. In: International Conference on Machine Learning. pp. 913–922. PMLR (2017)
9. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv. **40**(2) (May 2008). https://doi.org/10.1145/1348246.1348248, https://doi.org/10.1145/1348246.1348248
10. Dizaji, K.G., Zheng, F., Sadoughi, N., Yang, Y., Deng, C., Huang, H.: Unsupervised deep generative adversarial hashing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3664–3673 (2018)
11. Duan, Y., Lu, J., Wang, Z., Feng, J., Zhou, J.: Learning deep binary descriptor with multi-quantization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1183–1192 (2017)
12. Ge, T., He, K., Ke, Q., Sun, J.: Optimized product quantization for approximate nearest neighbor search. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2946–2953 (2013)

13. Gong, Y., Lazebnik, S., Gordo, A., Perronnin, F.: Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. TPAMI **35**(12), 2916–2929 (2012)
14. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
15. Hadi Kiapour, M., Han, X., Lazebnik, S., Berg, A.C., Berg, T.L.: Where to buy it: Matching street clothing photos in online shops. In: Proceedings of the IEEE international conference on computer vision. pp. 3343–3351 (2015)
16. Haseyama, M., Matsumura, A.: A trainable retrieval system for cartoon character images. In: 2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698). vol. 2, pp. II–393. IEEE (2003)
17. Heo, J.P., Lee, Y., He, J., Chang, S.F., Yoon, S.E.: Spherical hashing. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 2957–2964. IEEE (2012)
18. Huang, S., Xiong, Y., Zhang, Y., Wang, J.: Unsupervised triplet hashing for fast image retrieval. In: Proceedings of the on Thematic Workshops of ACM Multimedia 2017. pp. 84–92 (2017)
19. Huiskes, M.J., Lew, M.S.: The mir flickr retrieval evaluation. In: Proceedings of the 1st ACM international conference on Multimedia information retrieval. pp. 39–43 (2008)
20. Jang, Y.K., Cho, N.I.: Self-supervised product quantization for deep unsupervised image retrieval. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12085–12094 (2021)
21. Jegou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. IEEE transactions on pattern analysis and machine intelligence **33**(1), 117–128 (2010)
22. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
23. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
24. Le, T.N.H., Yao, S.Y., Wu, C.T., Lee, T.Y.: Regenerating arbitrary video sequences with distillation path-finding. IEEE Transactions on Visualization and Computer Graphics pp. 1–14 (2023). https://doi.org/10.1109/TVCG.2023.3237739
25. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **2**(1), 1–19 (2006)
26. Li, X., Wang, G., et al.: Deep supervised discrete hashing. Advances in Neural Information Processing Systems (2017)
27. Li, Y., van Gemert, J.: Deep unsupervised image hashing by maximizing bit entropy (2020), https://arxiv.org/abs/2012.12334
28. Li, Y., van Gemert, J.: Deep unsupervised image hashing by maximizing bit entropy. In: AAAI. pp. 2002–2010 (2021)
29. Lin, K., Lu, J., Chen, C.S., Zhou, J.: Learning compact binary descriptors with unsupervised deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1183–1192 (2016)
30. Morozov, S., Babenko, A.: Unsupervised neural quantization for compressed-domain similarity search. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3036–3045 (2019)

31. Nickel, M., Kiela, D.: Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In: International conference on machine learning. pp. 3779–3788. PMLR (2018)
32. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
33. Qiu, Z., Liu, J., Chen, Y., King, I.: Hihpq: Hierarchical hyperbolic product quantization for unsupervised image retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4614–4622 (2024)
34. Qiu, Z., Su, Q., Ou, Z., Yu, J., Chen, C.: Unsupervised hashing with contrastive information bottleneck. In: IJCAI. pp. 959–965 (2021)
35. Salakhutdinov, R., Hinton, G.: Semantic hashing. International Journal of Approximate Reasoning **50**(7), 969–978 (2009)
36. Shen, Y., Liu, L., Shao, L.: Unsupervised binary representation learning with deep variational networks. International Journal of Computer Vision **127**(11), 1614–1628 (2019)
37. Shen, Y., Qin, J., Chen, J., Yu, M., Liu, L., Zhu, F., Shen, F., Shao, L.: Auto-encoding twin-bottleneck hashing. In: CVPR. pp. 2818–2827 (2020)
38. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
39. Song, J., He, T., Gao, L., Xu, X., Hanjalic, A., Shen, H.T.: Binary generative adversarial networks for image retrieval. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
40. Su, S., Zhang, C., Han, K., Tian, Y.: Greedy hash: Towards fast optimization for accurate hash coding in cnn. In: NeurIPS (2018)
41. Toon Boom Animation Inc.: Toon boom harmony. Computer software (2023), https://www.toonboom.com/products/harmony, version 21
42. Wang, J., Zeng, Z., Chen, B., Dai, T., Xia, S.T.: Contrastive quantization with code memory for unsupervised image retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2468–2476 (2022)
43. Weiss, Y., Torralba, A., Fergus, R.: Spectral hashing. In: NeurIPS (2008)
44. Yang, E., Deng, C., Liu, T., Liu, W., Tao, D.: Semantic structure-based unsupervised deep hashing. In: Proceedings of the 27th international joint conference on artificial intelligence. pp. 1064–1070 (2018)
45. Yang, Y., Zhuang, Y., Xu, D., Pan, Y., Tao, D., Maybank, S.: Retrieval based interactive cartoon synthesis via unsupervised bi-distance metric learning. In: Proceedings of the 17th ACM international conference on multimedia. pp. 311–320 (2009)
46. Yu, T., Meng, J., Fang, C., Jin, H., Yuan, J.: Product quantization network for fast visual search. International Journal of Computer Vision **128**(8), 2325–2343 (2020)
47. Zhang, C., Zhu, X., Sun, A.: A flexible and scalable framework for video moment search. arXiv preprint arXiv:2501.05072 (2025)
48. Zieba, M., Semberecki, P., El-Gaaly, T., Trzcinski, T.: Bingan: Learning compact binary descriptors with a regularized gan. Advances in neural information processing systems **31** (2018)