Contrastive Transformer Masked Image Hashing for Degraded Image Retrieval

Xiaobo Shen[1] , **Haoyu Cai**[1] , **Xiuwen Gong**[2] and **Yuhui Zheng**[3*]
[1]Nanjing University of Science and Technology
[2]University of Technology Sydney
[3]Qinghai Normal University
njust.shenxiaobo@gmail.com, hycai@njust.edu.cn, gongxiuwen@gmail.com, zhengyh@vip.126.com

Abstract

Hashing utilizes hash code as a compact image representation, offering excellent performance in large-scale image retrieval due to its computational and storage advantages. However, the prevalence of degraded images on social media platforms, resulting from imperfections in the image capture process, poses new challenges for conventional image retrieval methods. To address this issue, we propose Contrastive Transformer Masked Image Hashing (CTMIH), a novel deep unsupervised hashing method specifically designed for degraded image retrieval, which is challenging yet relatively less studied. CTMIH addresses the problem by training on transformed and masked images, aiming to learn transform-invariant hash code in an unsupervised manner to mitigate performance degradation caused by image deterioration. CTMIH utilizes Vision Transformer (ViT) architecture applied to image patches to capture distant semantic relevance. CTMIH introduces cross-view debiased contrastive loss to align hash tokens from augmented views of the same image and presents semantic mask reconstruction loss at the patch level to recover masked patch tokens. Extensive empirical studies conducted on three benchmark datasets demonstrate the superiority of the proposed CTMIH over the state-of-the-art in both degraded and normal image retrieval.

1 Introduction

Hashing [Wang *et al.*, 2017] has been widely applied for efficient retrieval from large-scale image databases due to its superiority of low computation and storage costs. It aims to convert high-dimensional image features into low-dimensional compact hashing codes while preserving similarity structure among images. Many learning-based hashing methods [Weiss *et al.*, 2008; Gong *et al.*, 2012; Wang *et al.*, 2017] have been proposed by employing machine learning on hash code generation, and are still being actively studied to support fast and accurate large-scale image retrieval.

*Corresponding author.

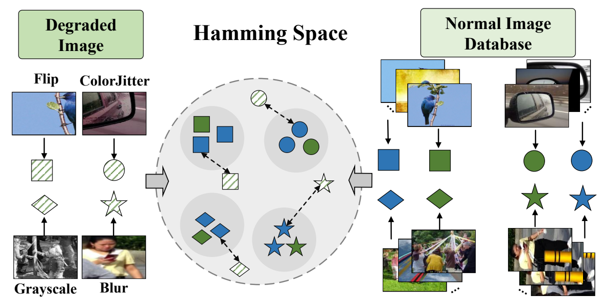


Figure 1: The illustration of degraded image retrieval. The four example images with different degradations, i.e., flip, color jitter, grayscale, and blur, are shown. In degraded image retrieval, the goal is to retrieve similar images from a normal image database given degraded image queries. To accomplish this, hash codes are utilized to represent both the degraded image queries (represented by solid shapes) and the normal image database (represented by hollow shapes). The retrieval process is performed in Hamming space. Image degradation often leads to discrepancies between the hash codes of degraded images and their original normal images. These discrepancies can significantly reduce retrieval performance.

There has been a recent research focus on deep hashing [Luo *et al.*, 2023] that introduces deep learning [LeCun *et al.*, 2015] into hashing, and learns image features and hash code in an end-to-end manner. With powerful learning capability of deep architectures, e.g., CNN [Krizhevsky *et al.*, 2012], ViT [Dosovitskiy *et al.*, 2021], deep hashing has shown its significant superiority over conventional shallow hashing [Weiss *et al.*, 2008; Gong *et al.*, 2012; Shen *et al.*, 2015]. Supervised deep hashing [Cao *et al.*, 2017; Li *et al.*, 2017; Liu *et al.*, 2019; Fan *et al.*, 2020] employs pairwise semantics or class label information to supervise training, and has achieved promising performance in image retrieval. However, supervised deep hashing heavily requires manually annotated labels, which are expensive to collect. Unsupervised deep hashing without need of semantic labels is valuable yet challenging in real applications. Until now some efforts [Shen *et al.*, 2020; Luo *et al.*, 2021; Qiu *et al.*, 2021; Ma *et al.*, 2022; Yu *et al.*, 2023] have been made towards unsupervised deep hashing.

Degraded images [Wang *et al.*, 2020; Yang *et al.*, 2022; Park *et al.*, 2023] are commonly encountered in various sce-

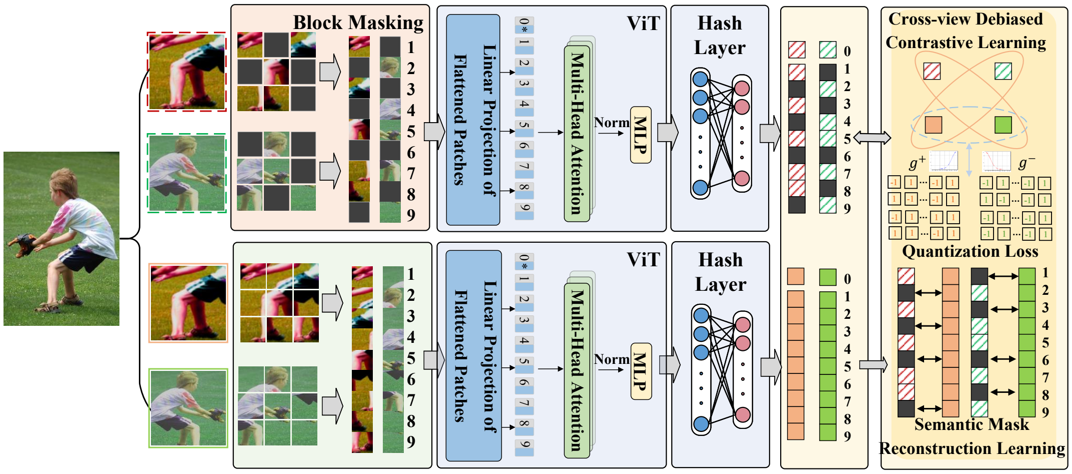


Figure 2: Illustration of the proposed CTMIH. CTMIH first generates two augmented views of an image by applying image transformation as image degradations, based on which two masked views are then obtained via the block masking strategy. The architecture of CTMIH consists of a weight-sharing Siamese network structure, where two branches receive patches from augmented views and masked views. Each branch employs a Vision Transformer (ViT) encoder and a hash layer to extract patch tokens and hash token, and quantizes hash token to obtain hash code. CTMIH is trained using the idea of self-distillation, minimizing three losses. The cross-view constrastive loss $\mathcal{J}_C$ aims to preserve the cross-view similarity between masked and augmented hash tokens, ensuring consistency across different views. The semantic mask reconstruction loss $\mathcal{J}_R$ aims to recover masked patch tokens to their corresponding augmented patch tokens, enhancing the quality of reconstructed images. The quantization loss $\mathcal{J}_Q$ reduces quantization error by treating it as a binary classification problem.

narios due to imperfections in the image capture process. These imperfections can result in serious degradations such as flipping, color jitter, grayscale, and blur, as illustrated in Figure 1. In practical applications such as pedestrian monitoring using surveillance cameras, it is often necessary to use the degraded pedestrian images as queries to retrieve similar high-quality images from a large-scale face image database [Li *et al.*, 2019]. The need for large-scale degraded image retrieval is evident, however this task is challenging and has received relatively less attention. Image degradation significantly affects semantic similarity of images, and may mislead training of conventional hashing, leading to reduced retrieval performance. Therefore, it is imperative to develop new deep hashing for degraded image retrieval, which remains relatively unexplored and presents substantial challenges.

To address this issue, we propose Contrastive Transformer Masked Image Hashing (CTMIH) that is specifically designed for degraded image retrieval. The key idea of CTMIH is to train hashing model on both transformed images and masked images to learn transform-invariant hash codes in an unsupervised manner, such that retrieval performance degradation caused by image degradation can be mitigated. The overview of the proposed CTMIH is illustrated in Figure 2. The main contributions of this work can be summarised as follows:

- We address the challenging task of degraded image retrieval, which has received relatively less attention. To tackle this, we propose a new deep unsupervised hashing method called Contrastive Transformer Masked Image Hashing (CTMIH).
- CTMIH utilizes ViT to encode image patches and learns transform-invariant hash code by aligning augmented and masked hash tokens while recovering masked patch tokens.
- Extensive empirical evaluations conducted on three benchmark image datasets demonstrate the superior performance of the proposed method over the state-of-the-art in both degraded and normal image retrieval.

2 Related Work

Deep Hashing Due to the strong learning capability of advanced deep architectures, e.g., CNN [Krizhevsky *et al.*, 2012], ViT [Dosovitskiy *et al.*, 2021], deep hashing [Luo *et al.*, 2023] has achieved promising performance in large-scale image retrieval. According to whether semantic supervision is used or not, deep hashing methods can be roughly divided into supervised deep hashing and unsupervised deep hashing. Supervised deep hashing[Cao *et al.*, 2017; Li *et al.*, 2017; Liu *et al.*, 2019; Fan *et al.*, 2020] typically outperforms unsupervised deep hashing by incorporating additional semantic labels. However, manual label annotation is time consuming and expensive, which is not often available in real applications.

Recently unsupervised deep hashing has been a hot research focus, as it overcomes the limitation of manual labeling. Some works [Dai *et al.*, 2017; Shen *et al.*, 2020] learn hash code by performing reconstruction tasks using some advanced deep architectures, e.g., variational autoencoders (VAE) [Kingma and Welling, 2013] and generative adversarial network (GAN) [Goodfellow *et al.*, 2014]. For instance, Semantic Structure based unsupervised Deep Hashing (SSDH) [Yang *et al.*, 2018] leverages two half Gaussian distributions to construct semantic structure, and further designs a pairwise similarity preservation loss. Bi-half

Net [Li and van Gemert, 2021] proposes to learn hash code by maximizing its bit entropy, and designs a parameter-free layer to force continuous image features to approximate the optimal half-half bit distribution. In addition, contrastive learning [Chen *et al.*, 2020; He *et al.*, 2020], as a powerful self-supervised method, has been introduced into deep unsupervised hashing, based on which some hashing methods have been proposed [Luo *et al.*, 2021; Qiu *et al.*, 2021; Ma *et al.*, 2022]. Contrastive Quantization with Code Memory (MeCoQ) [Wang *et al.*, 2022] proposes to use contrastive loss to better capture discriminative visual semantics and further use quantization code memory to enhance contrastive learning with lower feature drift. While many deep hashing methods employ CNN as a backbone, there have been a few recent works [Qiu *et al.*, 2021; Yu *et al.*, 2022; Ng *et al.*, 2023] that consider ViT as a backbone to encode images. Most existing deep unsupervised hashing methods have been developed for mainly conventional normal image retrieval, and may not perform well in degraded image retrieval, where image queries are degraded.

Degraded Image Analysis There have been a few attempts [Wang *et al.*, 2020; Yang *et al.*, 2022] on degraded image analysis, mainly on degraded image classification. Based on the observation that the distributions of corresponding patches in high- and low-quality images have uniform margins, feature de-drifting module (FDM) [Wang *et al.*, 2020] is proposed to learn the mapping relationship between deep representations of high- and low-quality images, and is further leveraged as a deep degradation prior to degraded image classification. In addition, a self-feature distillation method with uncertainty modeling [Yang *et al.*, 2022] is proposed for degraded image classification. It employs the high-quality features to distill the corresponding degraded ones and conduct uncertainty modeling, increasing importance of feature regions that are difficult to recover. However, to our knowledge, until now there have been few attempts of learning hash code for degraded image retrieval.

3 The Proposed Method

3.1 Problem Setup

Generally speaking, the goal of deep hashing for image retrieval is to map an ideal image $\mathbf{x}$ to hash code $\mathbf{b}$, which is used to support fast image retrieval in Hamming space. However, in real-world applications, due to various sources of degradation, e.g., noise, motion blur, and ColorJitter, we only observe the degraded image $\tilde{\mathbf{x}}$ instead of the ideal one. This work develops deep hashing for a challenging yet less studied task, i.e., degraded image retrieval, where image queries are degraded, while images in database are normal.

In degraded image retrieval, given a degraded image query $\tilde{\mathbf{x}}$, the aim of the proposed method is to learn a deep hash function $\mathcal{H} : \tilde{\mathbf{x}} \longrightarrow \mathbf{b} \in \{-1, 1\}^L$, where $\mathbf{b}$ is the hash code, L is code length. The learned hash code is then employed to achieve fast and accurate retrieval of the degraded image query from a normal image database. As the degraded images have different inherent appearances from normal images, it can result in obvious retrieval performance degradation using the conventional hash function. To address this issue, the key of the proposed method is to learn the transform-invariant hash function that is robust to image degradation. As such, the degraded image has similar hash code to its normal image, making it possible to quickly and accurately retrieve images that are visually or semantically similar to the degraded image query.

Algorithm 1 Image Transformation $\mathcal{T}$

Input: Image $\mathbf{X}$, hyper-parameter δ;
Output: Transformed Image.

1: Crop $\mathbf{X}$ with size of 256×256 randomly;
2: Resize $\mathbf{X}$ to size of 224×224;
3: Flip $\mathbf{X}$ horizontally with a probability of 0.5δ;
4: Add colorjitter to $\mathbf{X}$ with a probability of 0.8δ;
5: Convert $\mathbf{X}$ to grayscale image with a probability of 0.4δ;
6: Apply Gaussian blur to $\mathbf{X}$ with a probability of 0.5δ.

3.2 Formulation

Architecture The most common and natural solution for mitigating retrieval performance degradation caused by image degradation is to train a deep hashing model with augmented images having various image transformations. For a given image $\mathbf{X}$, the two random augmentations are applied, yielding two distorted views: $\mathbf{U} = \mathcal{T}(\mathbf{X}, \delta_u)$ and $\mathbf{V} = \mathcal{T}(\mathbf{X}, \delta_v)$, where $\mathcal{T}$ denotes the transformation function that performs image degradation, hyper-parameter δ_u and δ_v control the degrees of the two transformation. In this work, the transformations include random crop, resize, flip, color jitter, and blur, and the transformation procedure is illustrated in Algorithm 1. It is clear that large δ leads to heavy image degradation. In addition, motivated by the success of masked image modeling, the blockwise masking is applied to two views $\mathbf{U}$ and $\mathbf{V}$ to obtain masked views $\tilde{\mathbf{U}}$ and $\tilde{\mathbf{V}}$ respectively. Specifically, for instance, given $\mathbf{U}$, we first divide it into K non-overlapping patches $\mathbf{u}_k$, where $k = 1, \ldots, K$, and then, with a masking ratio r , perform masking on a random subset of patches, which are replaced by $\mathbf{e}$.

As illustrated in Figure 2, Siamese network structure with two branches sharing network weights is employed in CT-MIH, and the patches of augmented views and masked views are fed into the two branches. Existing deep hashing methods mainly employ CNN as the backbone that performs convolution operations on a small neighborhood of an image, and struggles to relate concepts spatially apart. Vision Transformer (ViT) [Dosovitskiy *et al.*, 2021] can effectively capture distant semantic relevance in an image by applying self-attention to a series of patches in an image. Inspired by the powerful capability of ViT on image modeling, we propose to employ ViT as an encoder in each branch, which receives image patches as sequential data and generates their latent representations. Following ViT encoder, a hash layer consists of a fully-connected layer and the sign function. For instance, given augmented view $\mathbf{U}$ and its masked view $\tilde{\mathbf{U}}$, hash layer outputs their hash tokens $\mathbf{h}^u$ and $\mathbf{h}^{\tilde{u}}$, and patch tokens $\mathbf{p}_k^u$ and $\mathbf{p}_k^{\tilde{u}}$, where $k = 1, \ldots, K$, and finally binarizes $\mathbf{h}^u$ to obtain hash code $\mathbf{b}^u = \text{sign}(\mathbf{h}^u)$.

Cross-view Debiased Contrastive Learning For conventional hashing methods, a degraded image tends to produce different hash code compared to its original counterpart due to changes in appearance, leading to unsatisfactory performance in degraded image retrieval. To mitigate this issue, a possible solution is to leverage contrastive learning (CL) [Chen *et al.*, 2020; He *et al.*, 2020] that aims to distinguish between semantically similar and dissimilar pairs of samples. Based on a triplet set that consists of anchor, positive, and negative samples, CL encourages the anchor and positive sample to be pulled closer together while pushing away the anchor from the negative samples. This work considers the benefit of CL that can well align the hash tokens of the two views augmented from one image, and push away the hash tokens of different images.

Specifically, given the i-th image $\mathbf{X}_i$, the hash tokens of its two augmented views are denoted as $\mathbf{h}_i^u$ and $\mathbf{h}_i^v$, and its two masked views are denoted as $\mathbf{h}_i^{\tilde{u}}$, $\mathbf{h}_i^{\tilde{v}}$. Taking $\mathbf{h}_i^u$ as anchor, we regard another masked view $\mathbf{h}_i^{\tilde{v}}$ as positive sample, and $\{\mathbf{h}_j^{\tilde{v}}\}_{j=1,j\neq i}^n$ as negative samples, where n is batch size. In this work, we contrast augmented and masked views with the purpose of distilling knowledge of unmasked views to help learn the hash token of masked views. Our idea is to enable the positive pairs to be close and the negative pairs to be apart. Treating $\mathbf{h}_i^u$ as anchor, we minimize the following conventional CL loss:

$$-\log \frac{o\left(\mathbf{h}_i^u, \mathbf{h}_i^{\tilde{v}}\right)}{o\left(\mathbf{h}_i^u, \mathbf{h}_i^{\tilde{v}}\right) + \sum_{j=1,j\neq i}^n o\left(\mathbf{h}_i^u, \mathbf{h}_j^{\tilde{v}}\right)} \quad (1)$$

where function $o\left(\mathbf{h}_i^u, \mathbf{h}_i^{\tilde{v}}\right) = \exp\left({\mathbf{h}^u_i}^\top \mathbf{h}_i^{\tilde{v}}/\tau\right)$ measures the similarity between two embeddings, τ is a temperature coefficient that controls dynamic range of product. The ℓ_2 normalization is applied on input embeddings before computing the inner product, such that the inner product is equivalent to cosine similarity.

As can be seen in (1), in conventional CL, negative points are randomly sampled from the whole set as true labels are unavailable. However, some of these negative samples may actually belong to the same class as anchor, referred as sampling bias, and this bias has been shown to degrade performance. To mitigate this gap, debiased CL is employed to correct sampling bias of negative samples, and its loss is defined as follows:

$$\mathcal{J}_C(\mathbf{h}_i^u, \mathbf{h}_i^{\tilde{v}}) = -\log \frac{o\left(\mathbf{h}_i^u, \mathbf{h}_i^{\tilde{v}}\right)}{o\left(\mathbf{h}_i^u, \mathbf{h}_i^{\tilde{v}}\right) + n\varphi\left(\mathbf{h}_i^u, \{\mathbf{h}_j^{\tilde{v}}\}_{j=1}^n\right)} \quad (2)$$

where the negative similarity term can be modified as:

$$\varphi\left(\mathbf{h}_i^u, \{\mathbf{h}_j^{\tilde{v}}\}_{j=1}^n\right) = \frac{1}{\varrho^-}\Big(\frac{1}{n}\sum_{j=1}^n o\left(\mathbf{h}_i^u, \mathbf{h}_j^{\tilde{v}}\right) - \varrho^+ o\left(\mathbf{h}_i^u, \mathbf{h}_i^{\tilde{v}}\right)\Big) \quad (3)$$

where ϱ^- and ϱ^+ represent the class probability that two images belong to same and different classes respectively, and we have $\varrho^- + \varrho^+ = 1$. Similarly, we can further define $\mathcal{J}_C(\mathbf{h}_i^v, \mathbf{h}_i^{\tilde{u}})$.

To this end, by considering all the n images, we have the following debiased CL loss:

$$\mathcal{J}_C = \frac{1}{n}\sum_{i=1}^n \Big(\mathcal{J}_C(\mathbf{h}_i^u, \mathbf{h}_i^{\tilde{v}}) + \mathcal{J}_C(\mathbf{h}_i^v, \mathbf{h}_i^{\tilde{u}})\Big) \quad (4)$$

Semantic Mask Reconstruction Learning Masked image patch reconstruction is a popular self-supervised pretext task with the idea of auto-encoding [Zhou *et al.*, 2021], and has been previously achieved by predicting raw pixels. This work leverages the benefits of masked image modeling to explore internal local structures in an image and better train ViT. Specifically, for instance, given one augmented view $\mathbf{U}_i$ and the corresponding masked view $\tilde{\mathbf{U}}_i$ of the i-th image, their patch token sequences are denoted as $\{\mathbf{p}_{ik}^u\}_{k=1}^K$ and $\left\{\mathbf{p}_{ik}^{\tilde{u}}\right\}_{k=1}^K$. In this work, we employ the idea of self-distillation, and instead of recovering raw pixels, propose to semantically recover the masked patch token using original patch token, i.e., enable the masked and its original tokens to be close. Specifically, given the i-th image, we recover its k-th image patch by minimizing the following cross-entropy (CE) loss:

$$\mathcal{J}_R\big(\mathbf{p}_{ik}^u, \mathbf{p}_{ik}^{\tilde{u}}\big) = -m_{ik}\mathbf{p}_{ik}^{u\top}\log \mathbf{p}_{ik}^{\tilde{u}} \quad (5)$$

where m_{ik} is set to 1 if the k-th patch of the i-th image is masked, and set to 0 otherwise. It simplifies masked recovering by turning it into a classification problem that is optimized by CE loss, and preserves more semantic information of patch token.

By considering all the image patches of all the n images, we have the following semantic mask reconstruction loss:

$$\mathcal{J}_R = \sum_{i=1}^n\sum_{k=1}^K \mathcal{J}_R\big(\mathbf{p}_{ik}^u, \mathbf{p}_{ik}^{\tilde{u}}\big) + \mathcal{J}_R\big(\mathbf{p}_{ik}^v, \mathbf{p}_{ik}^{\tilde{v}}\big) \quad (6)$$

3.3 Training and Inference

Training To obtain high-quality hash code, quantization loss is introduced to reduce quantization error. Motivated by the observation that hashing aims to predict the sign of each bit, this problem can be naturally regarded as the binary classification. Specifically, we employ a pre-defined Gaussian distribution estimator $g(h) = \exp(-\frac{(h-\mu)^2}{2\sigma^2})$ to evaluate binary likelihood of each hash bit, where μ and σ denote mean and standard deviation respectively. We then define $G(\cdot) = \{g^+, g^-\}$, where with the same σ, g^+ and g^- are defined with $\mu = 1$ and $\mu = -1$ respectively. The quantization loss is calculated as binary cross-entropy classification loss (BCE). Taking $\mathbf{h}_i^u$ for example, its quantization loss is defined as follows:

$$\mathcal{J}_Q(\mathbf{h}_i^u) = \frac{1}{L}\sum_{l=1}^L \Big(Q\left(y_l^+, g_l^+\right) + Q\left(y_l^-, g_l^-\right)\Big) \quad (7)$$

where BCE loss is defined as $Q(y, g) = -y\log g + (1 - y)\log(1-g)$, $g_l^+ = g^+(\mathbf{h}_{il}^u)$ and $g_l^- = g^-(\mathbf{h}_{il}^u)$ denote the two estimated likelihoods of l-th hash bit of $\mathbf{h}_i^u$, $y_l^+ = \frac{1}{2}(\text{sign}(\mathbf{h}_{il}^u) + 1)$ and $y_l^- = 1 - y_l^+$ denote the likelihood labels, L is the code length. In this way, the quantization error

Method	Refernce	Backbone	MS COCO			NUS-WIDE			ImageNet		
			16 bits	32 bits	64 bits	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
SH	NeurIPS 08	VGG	0.524	0.529	0.544	0.528	0.523	0.548	0.214	0.325	0.406
ITQ	TPAMI 13	VGG	0.598	0.646	0.667	0.612	0.618	0.646	0.327	0.461	0.549
SSDH	IJCAI 18	VGG	0.540	0.564	0.592	0.673	0.692	0.707	0.176	0.217	0.248
GreedyHash	NeurIPS 18	VGG	0.549	0.570	0.617	0.637	0.646	0.688	0.147	0.307	0.434
Bi-half Net	AAAI 21	VGG	0.638	0.711	0.734	0.736	0.756	0.766	0.439	0.537	0.591
MeCoQ	AAAI 22	VGG	0.692	0.738	0.753	0.744	0.776	0.788	0.599	0.646	0.710
OH	ACM MM 23	VGG	0.656	0.667	0.692	0.711	0.742	0.771	0.654	0.674	0.705
CIBHash	IJCAI 21	ViT	<u>0.767</u>	0.796	0.813	<u>0.781</u>	0.801	0.813	0.732	0.768	0.795
WCH	ACCV 22	ViT	0.741	0.775	0.794	<u>0.781</u>	<u>0.804</u>	<u>0.819</u>	0.733	0.775	0.796
SDC	BMVC 23	ViT	0.765	<u>0.801</u>	<u>0.823</u>	0.764	0.796	0.809	<u>0.734</u>	<u>0.779</u>	<u>0.814</u>
CTMIH	Ours	ViT	**0.809**	**0.834**	**0.846**	**0.795**	**0.816**	**0.826**	**0.820**	**0.860**	**0.869**

Table 1: mAPs of all the hashing methods for degraded image retrieval on the three benchmark datasets. The bold and underlined indicate best and second best, respectively.

is reduced by minimizing the above binary classification loss. Considering the hash tokens of all the n images, we have the following quantization loss:

$$\mathcal{J}_Q = \sum_{i=1}^{n} \mathcal{J}_Q(\mathbf{h}_i^u) + \mathcal{J}_Q(\mathbf{h}_i^v) \tag{8}$$

To this end, by summarizing the three losses, i.e., $\mathcal{J}_C$, $\mathcal{J}_R$, $\mathcal{J}_Q$, we have the following objective function of the proposed method:

$$\mathcal{J} = \mathcal{J}_C + \alpha\mathcal{J}_R + \beta\mathcal{J}_Q \tag{9}$$

where α and β are two non-negative parameters to balance each term.

Inference Once the proposed CTMIH is trained, and given an image query, we first divide it into several patches and further feed it into the ViT encoder and hash layer without masking to generate its hash code. For degraded image retrieval, the image query is required to be degraded. We perform transformations on a normal image to obtain a degraded image query, which is detailed in the experiment setup.

4 Experiments

4.1 Experimental Setup

Datasets The experiments are conducted on three benchmark image datasets, which are detailed as follows:

MSCOCO [Lin *et al.*, 2014] is a large-scale image dataset for object detection, segmentation, and captioning. Following previous works, a subset of 122,218 images from 80 categories is used, where the 5,000 images are randomly selected as the query set and the remaining images are used as the database. The 10,000 images are randomly selected from the database for training.

NUS-WIDE [Chua *et al.*, 2009] is a multi-label dataset consisting of 269,648 images from 81 categories. A subset with images from the 21 most frequent categories is used. The 100 images are randomly sampled from each category as the query set and the remaining images are used as the database. The 500 images for each category are randomly sampled from the database for training.

ImageNet [Russakovsky *et al.*, 2014] is a single-label image dataset, where each image is labeled by one of 1, 000 categories. A subset with 143,495 images in 100 categories is used, where 100 images from each category are randomly sampled for training, 5,000 images are sampled as the query set, and the remaining images are used as the database.

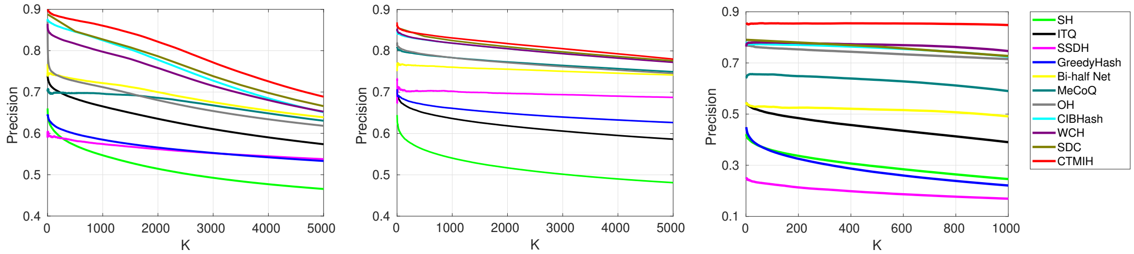
Baselines We compare the proposed method with various state-of-the-art unsupervised hashing baselines, including two shallow hashing methods, i.e., SH [Weiss *et al.*, 2008] and ITQ [Gong *et al.*, 2012], five CNN based deep hashing methods, i.e., SSDH [Yang *et al.*, 2018], GreedyHash [Su *et al.*, 2018], Bi-half Net [Li and van Gemert, 2021], MeCoQ [Wang *et al.*, 2022], OH [Yu *et al.*, 2023], three ViT based deep hashing methods, i.e., CIBHash [Qiu *et al.*, 2021], WCH [Yu *et al.*, 2022], SDC [Ng *et al.*, 2023].

Experimental Setting For all the methods, the images are resized to $224 \times 224 \times 3$. For shallow hashing methods, the 4,096-dimensional feature extracted by the VGG-F model pre-trained on ImageNet is used for training. For deep hashing methods, the raw image is directly used as for training. For the proposed method, we apply Algorithm 1 on each image in the training set to generate two augmented views, where δ_u and δ_v are set to 0.5 and 1 respectively. To generate the degraded image query, we apply Algorithm 1 on each image in the query set by setting δ to 0.5 by default. The standard ViT-Base is used as the backbone, and the size and number of patches are set to 16 and 196 respectively. The masking ratio r is set to 0.3, class probability ϱ^+ is set to 0.05, and temperature τ is set to 0.5. The two hyper-parameters α and β are set to 0.1 and 0.1 respectively. The batch size is set to 32, the number of epochs is set to 100, and the learning rates of ViT and hash layer are set to 10^{-5} and 10^{-3} respectively. The proposed method is trained using Adam optimizer.

Evaluation Metrics Following [Zhang *et al.*, 2017], we consider the widely-used mean Average Precision (mAP@K) and Precision curve as evaluation metrics, and K is set to 5000 for NUS-WIDE and MSCOCO, and 1000 for ImageNet.

4.2 Comparisons with State-of-the-art

Results on Degraded Image Retrieval This section evaluates the performance of degraded image retrieval. The mAPs of the proposed CTMIH and ten state-of-the-art hashing baselines on three benchmark datasets are reported in Table 1. In addition, precision curves of all the methods with respect to



(a) MS COCO (b) NUS-WIDE (c) ImageNet

Figure 3: Precision curves of all the hashing methods with respect to 32 bits on three benchmark datasets.

Method	MS COCO	NUS-WIDE	ImageNet
SH	0.585 ↘ 0.061	0.638 ↘ 0.110	0.338 ↘ 0.124
ITQ	0.658 ↘ 0.060	0.718 ↘ 0.102	0.439 ↘ 0.112
SSDH	0.556 ↘ 0.016	0.696 ↘ 0.023	0.229 ↘ 0.053
GreedyHash	0.586 ↘ 0.037	0.675 ↘ 0.038	0.231 ↘ 0.084
Bi-half Net	0.645 ↘ 0.007	0.750 ↘ 0.014	0.540 ↘ 0.101
MeCoQ	0.755 ↘ 0.063	0.770 ↘ 0.026	0.705 ↘ 0.051
OH	0.748 ↘ 0.092	0.796 ↘ 0.085	0.672 ↘ 0.018
CIBHash	0.793 ↘ 0.026	$\underline{0.790}$ ↘ 0.009	$\underline{0.783}$ ↘ 0.051
WCH	0.760 ↘ 0.019	0.788 ↘ 0.007	0.761 ↘ 0.028
SDC	$\underline{0.810}$ ↘ 0.045	0.787 ↘ 0.023	0.764 ↘ 0.030
CTMIH	**0.818** ↘ **0.009**	**0.799** ↘ **0.004**	**0.832** ↘ **0.012**

Table 2: mAPs of all the hashing methods for normal image retrieval on the three benchmark datasets. The mAP drops on degraded image retrieval compared to normal image retrieval are also reported.

32 bits are shown in Figure 3. From Table 1 and Figure 3, we can clearly observe that 1) the proposed CTMIH outperforms all the baselines in all the 10 cases. Specifically, it outperforms the baselines averagely by 3.3%, 1.1%, 7.4% on MS COCO, NUS-WIDE, ImageNet respectively. In addition, the precision curves of CTMIH are generally above those of the baselines. 2) among all the baselines, CIBHash, WCH, SDC with ViT backbone outperform the other deep hashing baselines with VGG backbone by a margin, followed by shallow hashing baselines. The empirical results clearly demonstrate the superiority of the proposed CTMIH for degraded image retrieval.

Results on Normal Image Retrieval This section evaluates the performance of conventional normal image retrieval, where the image queries are normal. The mAPs of all the hashing methods with respect to 16 bits are reported in Table 2. From this table, we see that the proposed CTMIH outperforms all the baselines, indicating that the proposed method also works well on conventional normal image retrieval. In addition, compared to normal image retrieval, the mAP drops of the proposed method on degraded image retrieval are lower than those of the baselines. This suggests that the proposed CTMIH is superior in mitigating the performance degradation caused by image degradation.

4.3 Further Analysis

Evaluation on Varying Degrees of Image Degradation The section evaluates the sensitivity of the deep hashing methods with respect to varying degrees of image degradation. We vary the parameter δ from the range of $[0, 0.2, 0.4, 0.6, 0.8, 1.0]$ to generate image queries with varying degrees of degradation, where the sample image query is shown in Figure 4. Figure 4 reports the mAPs of all the deep hashing methods with varying degrees of degradation on ImageNet. From this figure, we can clearly observe that mAP of the proposed CTMIH is not relatively insensitive to the change of δ, compared to the deep hashing baselines. Specifically, as the degree of image degradation increases, the mAP of the proposed CTMIH decreases by 1.3%, while the mAPs of the baselines have shown a decrease ranging from 5.2% to 20.1%. The above results clearly demonstrate that the proposed CTMIH is robust to image degradation, and performs well on degraded image retrieval.

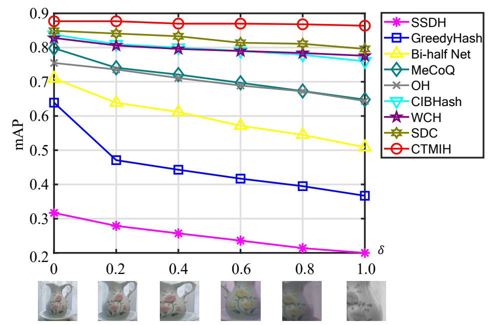
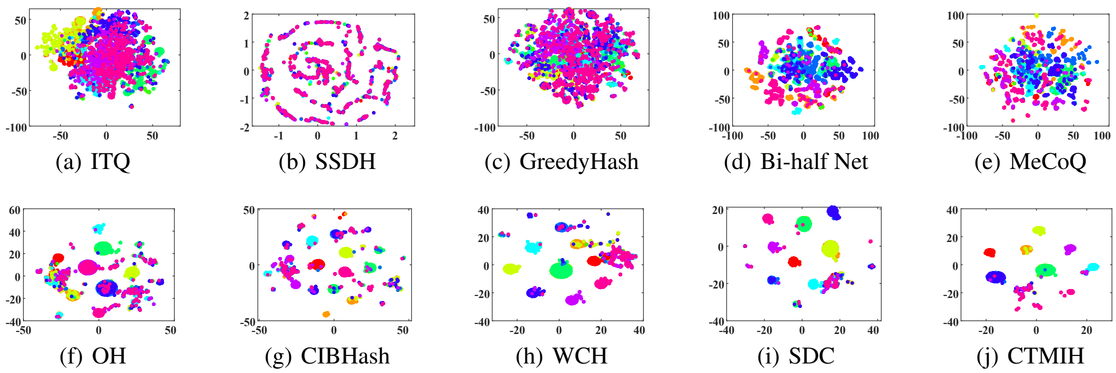


Figure 4: mAPs of all the deep hashing methods with respect to varying degrees of image degradation on ImageNet.

Method	MS COCO	NUS-WIDE	ImageNet
Naive CL	0.719	0.698	0.674
w/o $\mathcal{J}_Q$	0.765	0.731	0.757
w/o $\mathcal{J}_R$	0.792	0.779	0.806
w/o $\mathcal{J}_C$	0.428	0.535	0.413
CTMIH	**0.809**	**0.795**	**0.820**

Table 3: The mAPs of the proposed method and four variants on three datasets.

Ablation Study This section empirically evaluates the effectiveness of each loss in the proposed CTMIH. We compare the proposed method with a baseline that is trained with conventional contrastive loss without masking strategy, and its three variants without each of three losses (w/o). The mAPs



(a) ITQ (b) SSDH (c) GreedyHash (d) Bi-half Net (e) MeCoQ

(f) OH (g) CIBHash (h) WCH (i) SDC (j) CTMIH

Figure 5: The t-SNE visualization of ImageNet using ten hashing methods. The different colors indicate different labels and a total of 10 classes with 1,000 samples from each class are randomly sampled for visualization.

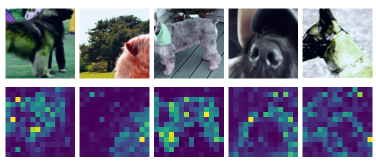


Figure 6: The visualization of attention maps of the proposed CTMIH on 5 randomly selected degraded images from ImageNet.

Figure 7: Top-10 retrieved results of the four deep hashing methods on a normal image and its degraded image queries from ImageNet.

of these methods with respect to 16 bits on the three datasets are reported in Table 3. From Table 3, we can clearly observe that removing $\mathcal{J}_Q$, $\mathcal{J}_R$, $\mathcal{J}_C$ results in an average decrease of 5.7%, 1.6%, 34.9% in mAP respectively on the three datasets, where the debiased contrastive loss has the greatest impact on performance. In addition, CTMIH obviously outperforms naive CL baseline by 11.10% on average.

Embedding Visualization This section qualitatively compares different hashing methods by visualizing learned hash codes of degraded images. We conduct experiment on ImageNet, randomly selecting 10,000 samples that belong to 10 classes, and setting code length to 32. The hash codes learned by ten methods are visualized into a 2-dimensional space with t-SNE [van der Maaten and Hinton, 2008], as illustrated in Figure 5. From Figure 5, we see that the visualization results are generally consistent with the quantitative empirical results.

Attention Map Visualization This section qualitatively evaluates the effectiveness of the proposed CTMIH by visualizing its attention maps. Figure 6 illustrates the attention maps generated by CTMIH for 5 randomly selected degraded images from ImageNet. From the figure, we can observe that the generated attention maps focus on the key components of images through masking reconstruction task. It indicates that the proposed CTMIH can effectively identify and prioritize important regions in the image.

Case Study This section presents a case study comparing the performance of the proposed CTMIH with three baselines, namely CIBHash, WCH, and SDC, for both degraded and normal image retrieval tasks. Figure 7 illustrates the top-10 retrieved images of one randomly selected normal image and its degraded image queries from ImageNet. A retrieved sample is marked in green if its label matches that of the query, and it is marked in red otherwise. From the results shown, we can observe that CTMIH outperforms the three baselines in terms of retrieving correct images in both tasks. These findings further confirm the effectiveness of the proposed CTMIH for both degraded and normal image retrieval tasks.

5 Conclusions

This work presents a preliminary attempt to learn hash code from degraded images, and proposes Contrastive Transformer Masked Image Hashing (CTMIH) for this challenging yet less studied degraded image retrieval. The proposed CTMIH aims to learn transform-invariant hash code through unsupervised training on transformed and masked images, mitigating performance degradation caused by image degradation. CTMIH aligns hash tokens of augmented views from the same image by performing constrastive learning, and recovers masked patch tokens with the idea of masked image prediction. The extensive empirical studies demonstrate the proposed CTMIH performs effectively on both degraded and normal image retrieval.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62176126, U20B2065, U22B2056, 62272468, the Natural Science Foundation of Jiangsu Province, China under Grant No. BK20230095, BK20211539.

References

[Cao *et al.*, 2017] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Philip S Yu. Hashnet: Deep learning to hash by continuation. In *Proceedings of IEEE International Conference on Computer Vision*, pages 5608–5617, 2017.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of International Conference on Machine Learning*, volume 119, pages 1597–1607, 2020.

[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 *Proceedings of ACM International Conference on Image and Video Retrieval*, pages 1–9, 2009.

[Dai *et al.*, 2017] Bo Dai, Ruiqi Guo, Sanjiv Kumar, Niao He, and Le Song. Stochastic generative hashing. In *Proceedings of International Conference on Machine Learning*, pages 913–922, 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, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of International Conference on Learning Representations*, 2021.

[Fan *et al.*, 2020] Lixin Fan, Kam Woh Ng, Ce Ju, Tianyu Zhang, and Chee Seng Chan. Deep polarized network for supervised learning of accurate binary hashing codes. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 825–831, 2020.

[Gong *et al.*, 2012] 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.

[Goodfellow *et al.*, 2014] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63:139 – 144, 2014.

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 9726–9735, 2020.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of International Conference on Learning Representations*, 2013.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1106–1114, 2012.

[LeCun *et al.*, 2015] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

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

[Li *et al.*, 2017] Qi Li, Zhenan Sun, Ran He, and Tieniu Tan. Deep supervised discrete hashing. In *Proceedings of Advances in Neural Information Processing Systems*, pages 2482–2491, 2017.

[Li *et al.*, 2019] Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *IEEE Transactions on Image Processing*, 28(4):1575–1590, 2019.

[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 *Proceedings of European Conference on Computer Vision*, pages 740–755, 2014.

[Liu *et al.*, 2019] Haomiao Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Deep supervised hashing for fast image retrieval. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 127(9):1217–1234, 2019.

[Luo *et al.*, 2021] 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 *Proceedings of International Joint Conference on Artificial Intelligence*, pages 902–908, 2021.

[Luo *et al.*, 2023] Xiao Luo, Haixin Wang, Daqing Wu, Chong Chen, Minghua Deng, Jianqiang Huang, and Xian-Sheng Hua. A survey on deep hashing methods. *ACM Transactions on Knowledge Discovery from Data*, 17(1):15:1–15:50, 2023.

[Ma *et al.*, 2022] Zeyu Ma, Xiao Luo, Yingjie Chen, Mixiao Hou, Jinxing Li, Minghua Deng, and Guangming Lu. Improved deep unsupervised hashing with fine-grained semantic similarity mining for multi-label image retrieval. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 1254–1260, 2022.

[Ng *et al.*, 2023] Kam Woh Ng, Xiatian Zhu, Jiun Tian Hoe, Chee Seng Chan, Tianyu Zhang, Yi-Zhe Song, Tao Xiang, Universiti Malaya CISiP, and GACRD Center. Unsupervised hashing with similarity distribution calibration. 2023.

[Park *et al.*, 2023] Dongwon Park, Byung Hyun Lee, and Se Young Chun. All-in-one image restoration for unknown degradations using adaptive discriminative filters for specific degradations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 5815–5824, 2023.

[Qiu *et al.*, 2021] Zexuan Qiu, Qinliang Su, Zijing Ou, Jianxing Yu, and Changyou Chen. Unsupervised hashing with contrastive information bottleneck. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 959–965, 2021.

[Russakovsky *et al.*, 2014] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115, 09 2014.

[Shen *et al.*, 2015] Fumin Shen, Chunhua Shen, Wei Liu, and Heng Tao Shen. Supervised discrete hashing. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 37–45, 2015.

[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 *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2827, 2020.

[Su *et al.*, 2018] Shupeng Su, Chao Zhang, Kai Han, and Yonghong Tian. Greedy hash: Towards fast optimization for accurate hash coding in cnn. *Proceedings of Advances in Neural Information Processing Systems*, 31, 2018.

[van der Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *The Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[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.*, 2020] Yang Wang, Yang Cao, Zheng-Jun Zha, Jing Zhang, and Zhiwei Xiong. Deep degradation prior for low-quality image classification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 11049–11058, 2020.

[Wang *et al.*, 2022] Jinpeng Wang, Ziyun Zeng, Bin Chen, Tao Dai, and Shu-Tao Xia. Contrastive quantization with code memory for unsupervised image retrieval. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 2468–2476, 2022.

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

[Yang *et al.*, 2018] Erkun Yang, Cheng Deng, Tongliang Liu, Wei Liu, and Dacheng Tao. Semantic structure-based unsupervised deep hashing. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 1064–1070, 2018.

[Yang *et al.*, 2022] Zhou Yang, Weisheng Dong, Xin Li, Jinjian Wu, Leida Li, and Guangming Shi. Self-feature distillation with uncertainty modeling for degraded image recognition. In *Proceedings of European Conference on Computer Vision*, pages 552–569, 2022.

[Yu *et al.*, 2022] Jiaguo Yu, Huming Qiu, Dubing Chen, and Haofeng Zhang. Weighted contrative hashing. In *Proceedings of the Asian Conference on Computer Vision*, pages 3861–3876, 2022.

[Yu *et al.*, 2023] Jiaguo Yu, Yuming Shen, and Haofeng Zhang. Hashing one with all. In *Proceedings of ACM International Conference on Multimedia*, pages 6420–6431, 2023.

[Zhang *et al.*, 2017] Haofeng Zhang, Li Liu, Yang Long, and Ling Shao. Unsupervised deep hashing with pseudo labels for scalable image retrieval. *IEEE Transactions on Image Processing*, 27(4):1626–1638, 2017.

[Zhou *et al.*, 2021] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *Proceedings of International Conference on Learning Representations*, 2021.