Let All Be Whitened: Multi-Teacher Distillation for Efficient Visual Retrieval

Zhe Ma[1], Jianfeng Dong[2,4*], Shouling Ji[1], Zhenguang Liu[1*], Xuhong Zhang[1], Zonghui Wang[1*], Sifeng He[3], Feng Qian[3], Xiaobo Zhang[3], Lei Yang[3]

[1]Zhejiang University,
[2]Zhejiang Gongshang University,
[3]Ant Group,
[4]Zhejiang Key Lab of E-Commerce
{mz.rs, sji, zhangxuhong, zhwang}@zju.edu.cn, {dongjf24, liuzhenguang2008, hsf215kg}@gmail.com,
{youzhi.qf, yl149505}@antgroup.com, ayou.zxb@antfin.com

Abstract

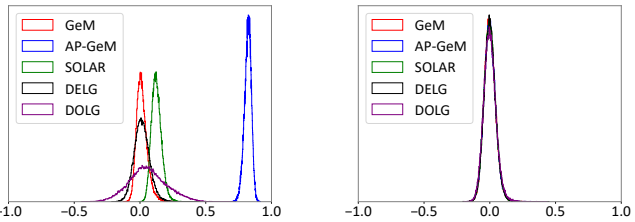
Visual retrieval aims to search for the most relevant visual items, e.g., images and videos, from a candidate gallery with a given query item. Accuracy and efficiency are two competing objectives in retrieval tasks. Instead of crafting a new method pursuing further improvement on accuracy, in this paper we propose a multi-teacher distillation framework Whiten-MTD, which is able to transfer knowledge from off-the-shelf pre-trained retrieval models to a lightweight student model for efficient visual retrieval. Furthermore, we discover that the similarities obtained by different retrieval models are diversified and incommensurable, which makes it challenging to jointly distill knowledge from multiple models. Therefore, we propose to whiten the output of teacher models before fusion, which enables effective multi-teacher distillation for retrieval models. Whiten-MTD is conceptually simple and practically effective. Extensive experiments on two landmark image retrieval datasets and one video retrieval dataset demonstrate the effectiveness of our proposed method, and its good balance of retrieval performance and efficiency. Our source code is released at https://github.com/Maryeon/whiten_mtd.

Introduction

Visual retrieval, such as image retrieval and video retrieval, is a long-standing problem in the computer vision community, which aims to search for the most similar items to a given query from a large number of candidates (Radenović, Tolias, and Chum 2018; Revaud et al. 2019; Noh et al. 2017; Yang et al. 2021). It supports a wide range of applications including instance matching (Cao, Araujo, and Sim 2020; Radenović, Tolias, and Chum 2018; Revaud et al. 2019), fine-grained recognition (Dong et al. 2021), product recommendation (Kim et al. 2021; Dong et al. 2018), etc.

Recent visual retrieval methods have been dominated by the Deep Neural Networks (DNNs) (He et al. 2016; Vaswani et al. 2017) based solutions. Given an overview of recent developments on visual retrieval problems (Chen et al. 2023; Shen, Hong, and Hao 2020), there are two considerations that must be taken into account. On the one hand, with the large amounts of pre-trained retrieval models being publicly available, it motivates us to make a complementary integration of them, in other words, take their essence and discard the dregs. On the other hand, despite the significant performance improvement boosted by DNNs, a major concern is their heavy overhead of computation. A compact and efficient retrieval model is required for large-scale scenarios. In this work, we jointly take the two considerations by aggregating multiple cumbersome pre-trained retrieval models into a more efficient one. To this end, we resort to knowledge distillation (Hinton et al. 2014; Passalis and Tefas 2018; Beyer et al. 2022), which is a technique proposed to transfer the knowledge from a large model or an ensemble of large models (teachers) to a small model (student) to reduce computation overhead, without significant loss in performance. We propose a multi-teacher distillation framework for efficient visual retrieval, where off-the-shelf pre-trained retrieval models are regarded as teachers and distilled to a more lightweight student model.

The keys to multi-teacher distillation are the *knowledge* to be transferred from teachers to a student, and the *fusion strategy* to combine the knowledge from teachers. The knowledge used in existing works can be categorized into feature-based (Hinton et al. 2014; Romero et al. 2015; Beyer et al. 2022) and relation-based (Yim et al. 2017; Tian, Krishnan, and Isola 2019; Park et al. 2019; Passalis,



(a) Without whitening. (b) With whitening.

Figure 1: Cosine similarity distributions of five existing landmark image retrieval models on GLDv2-clean dataset. Their similarity distributions are clearly different, making the similarities between different models incommensurable. Such discrepancy can be alleviated through whitening their output features.

*Corresponding authors.
Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Tzelepi, and Tefas 2020). Feature-based knowledge distillation constrains the student's intermediate or final outputs to be aligned to those of teacher models, forming a point-to-point fitting. Relation-based knowledge distillation typically transfers the similarity between layers (Yim et al. 2017; Passalis, Tzelepi, and Tefas 2020) or samples (Park et al. 2019; Tian, Krishnan, and Isola 2019) to the student model. As retrieval is generally modeled as a ranking problem where similarity scores defined between samples are used to sort the candidates, we employ the similarity between samples as the knowledge to be transferred.

Another key problem of multi-teacher distillation is how to fuse the knowledge of teacher models. Previous multi-teacher distillation works are mainly developed for classification models (Buciluǎ, Caruana, and Niculescu-Mizil 2006; Hinton et al. 2014; Shen, He, and Xue 2019; Yuan et al. 2021), whose knowledge is conveyed by the output categorical probabilistic distributions that has equivalent measure and are commensurable between different models. Hence, simple strategies like average (Buciluǎ, Caruana, and Niculescu-Mizil 2006; Hinton et al. 2014), random selection (Shen, He, and Xue 2019) or weighted summation (Yuan et al. 2021) can be reasonably utilized. However, due to the nature of retrieval methods that impose relational constraints on samples rather than push them to a finite number of prototypes (weights of the last classification layer) as in normal classification models, the representation space obtained by different visual retrieval models will be diversified and not commensurable. Figure 1a shows the cosine similarity distribution of five existing landmark image retrieval models, demonstrating the clear discrepancy between retrieval models. Therefore, simply and directly utilizing the fusion strategy defined for image classification is suboptimal for visual retrieval. How to fuse the knowledge of multiple teachers becomes a unique problem for retrieval tasks.

To alleviate the discrepancy of retrieval models, we propose to firstly align the similarity distributions of different retrieval models before any fusion operations. To this end, we propose to utilize whitening, which transforms teacher models' output into a spherical distribution. We will show that under spherical distribution, similarity distributions of teacher models can be perfectly aligned and more commensurable (as exemplified in Figure 1b), without sacrificing their distillation performance. Additionally, we devise multiple heuristic fusion strategies to aggregate predictions of teacher models. The fusion strategies compare and integrate the pairwise similarities calculated by each teacher model and produce the final knowledge to be transferred to the student model.

Our key contributions can be summarized as follows:

- We propose a simple but effective **Whiten**ing-based **M**ulti-**T**eacher **D**istillation (Whiten-MTD) framework for visual retrieval. Whiten-MTD is able to readily transfer knowledge from multiple strong and cumbersome retrieval models into a lightweight one. To the best of our knowledge, this is the first multi-teacher distillation work for visual retrieval.
- In order to make different retrieval models commensurable, we propose to whiten the output of teacher models to align the discrepant similarity distribution. Besides, we devise five heuristic fusion strategies, and demonstrate which one is best suited for our multi-teacher distillation framework by empirical study.
- We conduct extensive experiments on two kinds of visual retrieval tasks, *i.e.,* instance image retrieval and video retrieval. Compared to the state-of-the-art (Song et al. 2023; Lee et al. 2022; Kordopatis-Zilos et al. 2022) that are based on ResNet-50, ResNet-101 (He et al. 2016) or ViT-B (Dosovitskiy et al. 2021) and using the reranking strategy, we are able to achieve comparable performance without the reranking using more lightweight ResNet-18 or ResNet-34.

Related Work

Visual Retrieval

We review two specific fields of visual retrieval, *i.e.,* instance image retrieval, and near-duplicate video retrieval.

The goal of instance image retrieval is to search for images containing the same instance (*e.g.,* landmark, person) as the query. Earlier efforts are based on hand-crafted features while recent works (Dusmanu et al. 2019; Noh et al. 2017; Luo et al. 2019) tend to utilize DNN features. For DNN features, local features encode rich regional information and are usually used to perform geometric verification (Fischler and Bolles 1981). However, the local feature matching is time-consuming and always adopted in the reranking process. Global features summarize the images and are more compact. They are obtained by directly taking the activations after fully connected layer or pooling the feature maps after convolutional layer. Various pooling operators such as R-MAC (Tolias, Sicre, and Jégou 2015), GeM (Radenović, Tolias, and Chum 2018), are proposed to capture regional discriminative information. In addition, there are other works focusing on fine-tuning pre-trained classification models for instance image retrieval, including designing loss functions (Revaud et al. 2019), sampling strategies (Arandjelovic et al. 2016; Gordo et al. 2016), etc.

Near-duplicate video retrieval (Shen, Hong, and Hao 2020; Liu et al. 2013) focuses on video matching to construct robust video representations against severe transformations. Recent methods (He et al. 2021) all resort to DNNs. Based on pre-trained models to extract frame features, they focus on designing upper frame feature aggregation schemes, such as Bag-of-Words (Kordopatis-Zilos et al. 2017a), Deep Metric Learning (Kordopatis-Zilos et al. 2017b), Transformer (Shao et al. 2021), *etc.* Computation efficiency is much more crucial for videos, which usually contain tens of thousands of images.

Current state-of-the-art visual retrieval methods achieve their superior performance with deep neural networks, such as ResNet-101, pursuing the maximization of accuracy. We instead focus on the efficiency of the underlying backbones.

Distillation for Visual Retrieval

Knowledge distillation is a technique that was initially designed to improve model efficiency (Buciluǎ, Caruana, and

Niculescu-Mizil 2006; Hinton et al. 2014). Knowledge distillation has been widely explored in representation learning (Caron et al. 2021; Noroozi et al. 2018), object detection (Zhixing et al. 2021; Kang et al. 2021; Guo, Alvarez, and Salzmann 2021), action recognition (Liu et al. 2023), etc.

For visual retrieval which depends on similarity between samples, relation-based knowledge distillation methods can be felicitously exploited (Passalis and Tefas 2018; Fang et al. 2021; Dong et al. 2023). What's more, a specific framework called asymmetric retrieval has also been proposed to mitigate the accuracy-efficient trade-off by processing database samples with a large model while processing online queries with a lightweight one. Approaches (Wu et al. 2022; Duggal et al. 2021) in this framework typically distill the large model into a lightweight one and deploy it online.

Existing work for visual retrieval belongs to a single-teacher distillation setting while multi-teacher distillation is mainly studied in the context of image classification (Buciluǎ, Caruana, and Niculescu-Mizil 2006; Ba and Caruana 2014; You et al. 2017; Shen, He, and Xue 2019; Yuan et al. 2021). Fusion strategies of teacher models are the central topic of this problem. Averaging of all teacher models' predictions is a normal choice (Buciluǎ, Caruana, and Niculescu-Mizil 2006; Ba and Caruana 2014). Other strategies include voting (You et al. 2017), random selection (Shen, He, and Xue 2019), selection by reinforcement learning agent (Yuan et al. 2021), *etc.* .

Overall, the problem of multi-teacher distillation for visual retrieval is not well addressed and the intractable discrepancy between visual retrieval models hinders the direct usage of existing multi-teacher distillation methods.

Method

We expect to transfer the knowledge from multiple teacher models, typically strong but cumbersome, to a lightweight student model. After the distillation, the teacher models are discarded, and only the student model is retained for efficient visual retrieval. Without loss of generality, we formulate our method in the context of image retrieval, which can be easily adapted to other visual retrieval tasks.

In what follows, we first describe the basic similarity-based single-teacher distillation. Then we extend it to multi-teacher distillation based on aggregating similarity matrices output by different teacher models and introduce five heuristic similarity fusion strategies. Finally we propose to use whitening to eliminate the similarity discrepancy among teacher models, which enables more effective and stable multi-teacher distillation.

Single-Teacher Distillation

Existing single-teacher distillation works (Passalis and Tefas 2018; Fang et al. 2021; Wu et al. 2022) are analogous but only differ in the exact similarities used for distillation. For a clear validation of our multi-teacher distillation approach, we adopt the basic distillation method, although some advanced components can also be included, such as a queue structure (Fang et al. 2021).

Specifically, both teacher and student models are image encoders that embed images from a high-dimensional input space $\mathbb{R}^m$ into a low-dimensional representation space. We denote the teacher model as $\psi(\cdot) : \mathbb{R}^m \rightarrow \mathbb{R}^{n_t}$, and the student model as $\varphi(\cdot) : \mathbb{R}^m \rightarrow \mathbb{R}^{n_s}$, where n_t and n_s indicate the output dimension of the teacher model and student model respectively. Suppose $\psi(\cdot)$ and $\varphi(\cdot)$ are both l_2-normalized functions, *i.e.,* $\|\psi(\cdot)\|_2 = 1$ and $\|\varphi(\cdot)\|_2 = 1$. Let $\mathcal{B} = \{(x_i, y_i)\}_{i=1}^N$ be a batch of relevant positive image pairs. Accordingly, x_i and y_j $(i \neq j)$ are negative, meaning that they are semantically irrelevant. The image pairs can be sampled either in a supervised or unsupervised way. For example, in landmark instance image retrieval, according to the annotation, two images containing the same landmark are regarded as a positive pair, otherwise a negative pair. For video retrieval, they are obtained by image augmentation as widely used in self-supervised learning (Grill et al. 2020; Caron et al. 2021; Chen, Xie, and He 2021).

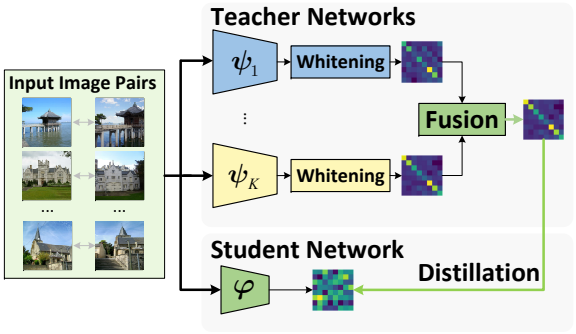


Figure 2: An overview of our multi-teacher distillation framework. Due to the similarity discrepancy of different retrieval models, the whitening is utilized on each teacher model to make their similarity commensurable. The fused similarity map as the knowledge of multiple teacher models is transferred to a lightweight student model. After distillation, the teacher models are discarded, and only the student model is retained for efficient visual retrieval.

We define cosine similarity as a measure of semantic relation between images, then a similarity matrix $\mathcal{T} \in \mathbb{R}^{N \times N}$ can be calculated using the teacher model ψ, where $\mathcal{T}[i,j] = \psi(x_i)^\top \psi(y_j)$. Similarly, $\mathcal{B}$ is represented by the student model as a similarity matrix $\mathcal{S} \in \mathbb{R}^{N \times N}$, where $\mathcal{S}[i,j] = \varphi(x_i)^\top \varphi(y_j)$. Moreover, we convert the similarity matrices $\mathcal{T}$ and $\mathcal{S}$ into probabilistic similarity matrices $\mathcal{Q}$ and $\mathcal{P}$ which enhances the relativity between similarity scores rather than their absolute values. Specifically, each row of the similarity matrices is normalized into a probabilistic distribution using Gaussian kernel with a temperature:

$$\mathcal{P}[i,j] = \frac{e^{\mathcal{S}[i,j]/\tau}}{\sum_{k=1}^N e^{\mathcal{S}[i,k]/\tau}}, \mathcal{Q}[i,j] = \frac{e^{\mathcal{T}[i,j]/\tau}}{\sum_{k=1}^N e^{\mathcal{T}[i,k]/\tau}} \quad (1)$$

where τ is the temperature factor. Finally the Kullback-Leibler (KL) divergence is minimized between each row of teacher probabilistic similarity matrix $\mathcal{Q}$ and student probabilistic similarity matrix $\mathcal{P}$ to conduct the distillation:

$$\mathcal{L} = \frac{1}{N} \sum_i^N KL\left(\mathcal{P}_i || \mathcal{Q}_i\right). \quad (2)$$

Multi-Teacher Distillation

We now extend to multi-teacher distillation by considering $K > 1$ teacher models $\psi_k(\cdot) \in \mathbb{R}^m \rightarrow \mathbb{R}^{n_{t_k}}$. To combine the knowledge of multiple teacher models, we propose to fuse K similarity matrices $\mathcal{T}_k$ produced by the teacher models into an aggregated similarity matrix $\mathcal{T}$ using a fusion strategy function $\mathcal{F}$,

$$\mathcal{T} = \mathcal{F}(\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_K). \tag{3}$$

Although elaborate strategies can be designed, we study five heuristic fusion strategies which are shown to be effective in our experiments. All the fusion strategies perform element-wise comparison and selection. For each location (i, j), we first include two normal strategies, which are:

- *mean*: $\mathcal{T}[i,j] = \frac{1}{K}\sum_{k=1}^{K} \mathcal{T}_k[i,j]$.
- *rand*: $\mathcal{T}[i,j] = \mathcal{T}_r[i,j], r \in [1, K]$.

Additionally, we design other three strategies based on decoupling the similarity matrix into inner-class similarities (diagonal) and inter-class similarities (off-diagonal), which are *max-min*, *max-mean*, and *max-rand*. All these strategies choose to take the maximum of similarities on the diagonal, favoring more compact inner-class clustering, while differing in the choice on the off-diagonal. For off-diagonal similarities (*i.e.,* $i \neq j$), the rules are:

- *max-min*: $\mathcal{T}[i,j] = \min_{k \in [1,K]} (\mathcal{T}_k[i,j])$.
- *max-mean*: $\mathcal{T}[i,j] = \frac{1}{K}\sum_{k=1}^{K} \mathcal{T}_k[i,j]$.
- *max-rand*: $\mathcal{T}[i,j] = \mathcal{T}_r[i,j], r \in [1, K]$.

However, these fusion strategies are not sufficient for effective multi-teacher distillation unless coupled with teacher model whitening, which will be described next.

Eliminating Teacher Discrepancy

The fusion strategy $\mathcal{F}$ in Eq. 3 requires to make comparison between different teachers and then perform a selection. However, due to the ranking property of retrieval, only the relative order of samples is considered regardless of the absolute value of similarities. This enables so much flexibility for models' outputs that models with similar retrieval performance may have diverse output similarity distributions. This can be demonstrated from a statistical visualization of the similarity distributions of different teacher models, as depicted in Figure 1a. If the distribution gap is neglected, existing normal multi-teacher fusion strategies, such as averaging, will be less effective as the output will be biased to the extreme distribution of a specific teacher model.

To bridge the gap among different teacher models, we propose to whiten their output representations before calculating similarity scores. Whitening is a linear transformation that transforms the representations into spherical distribution. Followed by l_2-normalization, the resulting representations will distribute uniformly on the unit sphere surface (Ermolov et al. 2021). As proved by Cai, Fan, and Jiang *et al.* (2013), the angle θ between two independent random vectors distributed uniformly on the unit sphere surface in $\mathbb{R}^n$ converges to a distribution with the probability density function $f(\theta) = \frac{1}{\sqrt{\pi}} \cdot \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} \cdot (\sin\theta)^{n-2}, \theta \in [0, \pi]$. Therefore, cosine similarities of different teacher models calculated with the whitened representations will follow the same distribution (see statistical results in Figure 1b). Whitening has been applied in recent works as a post-processing technique for further performance improvement in instance matching (Cao, Araujo, and Sim 2020; Revaud et al. 2019) or normalization technique to prevent representation collapse in self-supervised learning (Ermolov et al. 2021; Weng et al. 2022). We otherwise uncover the capability of whitening to align teachers' similarity distributions, which is crucial for multi-teacher distillation.

Specifically, we adopt PCA-whitening which is learned on the training set and fixed in the distillation stage. Given a training set $\mathcal{D} = \{x_i\}_{i=1}^{M}$ where M is the total number of samples, for the k-th teacher, the PCA-whitening $\mathcal{W}_k$ is a linear transformation W_k:

$$\mathcal{W}_k(\psi_k(\cdot)) = W_k\left(\psi_k(\cdot) - b_k\right), \tag{4}$$

where $W_k \in \mathbb{R}^{n_c \times n_{t_k}}$, $b_k \in \mathbb{R}^{n_{t_k}}$, n_c is the dimensionality of whitened representations. $b_k = \frac{1}{M}\sum_{i=1}^{M} \psi_k(x_i)$ is the mean representation of all the training samples and W_k is the eigen-system of Σ satisfying $W_k^\top W_k = \Sigma^{-1}$, where Σ is the covariance matrix of zero-meaned $\psi_k(\cdot)$ (*i.e.,* $\psi_k(\cdot) - b_k$). We provide a detailed discussion on the efficacy of whitening and the choice of n_c in supplementary materials.

Experiments

To verify the viability of Whiten-MTD, we evaluate it on two common visual retrieval tasks, *i.e.,* landmark image retrieval and near-duplicate video retrieval. We provide implementation details and additional experiment results in supplementary materials.

Landmark Image Retrieval

Datasets Following previous works (Lee et al. 2022; Song et al. 2023), we use the clean version of Google Landmark Dataset v2 (GLDv2-clean) (Weyand et al. 2020) as the training set and two additional independent datasets RParis6k ($\mathcal{R}$Par) and ROxford5k ($\mathcal{R}$Oxf) (Radenović et al. 2018) for evaluation. GLDv2-clean consists of around 1.6 million images of 80K landmark instances. $\mathcal{R}$Oxf and $\mathcal{R}$Par are the revisited version of Paris6k (Philbin et al. 2008) and Oxford5k (Philbin et al. 2007) building datasets. The revisited versions improved the reliability of ground truth and introduced three new protocols of varying difficulty, *i.e.,* Easy, Medium and Hard. Both $\mathcal{R}$Par and $\mathcal{R}$Oxf contains 70 queries, and 6,322 and 4,993 gallery images respectively. Additionally, 1M distractor images are included as the extra gallery images for large-scale evaluation on both datasets.

Metrics As done in (Radenović, Tolias, and Chum 2018; Revaud et al. 2019), we report mAP under the Medium and Hard evaluation protocols on $\mathcal{R}$Par and $\mathcal{R}$Oxf, as well as large-scale evaluation with the 1M distractors. For the model complexity, we report the model size in terms of the number of parameters and computation overhead during the inference. The computation overhead is measured by the number of GFLOPs when a model encodes a given image of size 1024×768.

Method	Whiten?	$\mathcal{R}$Oxf		$\mathcal{R}$Par	
		M	H	M	H
R101-GeM	✗	69.88	45.00	82.69	65.13
	✓	68.97	45.08	82.15	65.02
R101-AP-GeM	✗	69.17	44.08	80.44	62.08
	✓	70.37	45.93	81.02	63.29
R101-SOLAR	✗	71.14	46.63	83.04	66.24
	✓	68.52	43.65	81.82	64.39
R101-DELG	✗	84.43	66.69	91.97	82.87
	✓	85.28	68.52	92.08	83.26
R101-DOLG	✗	80.09	61.64	88.68	77.01
	✓	83.27	64.95	89.78	78.58

Table 1: Influence of whitening on the performance of teacher models. The capacities of teacher models are largely retained after whitening. M: Medium, H: Hard.

Strategy	Whiten?	$\mathcal{R}$Oxf		$\mathcal{R}$Par	
		M	H	M	H
mean	✗	69.62	44.01	81.62	63.64
rand		65.68	38.89	75.78	55.09
max-min		67.63	40.78	82.00	65.01
max-mean		67.09	40.07	83.66	66.93
max-rand		71.53	46.71	80.26	62.15
mean	✓	71.11	46.38	82.55	65.62
rand		71.01	46.39	82.06	64.01
max-min		**74.67**	**50.69**	**84.48**	**68.34**
max-mean		73.90	49.53	83.42	66.55
max-rand		72.53	48.10	82.60	65.46

Table 2: Performance comparison of different similarity fusion strategies in distillation from R101-GeM, R101-AP-GeM and R101-SOLAR to ResNet-18. The whitening is beneficial for multi-teacher distillation, and it also makes it less sensitive to the selection of fusion strategies. The max-min fusion strategy performs the best with whitening.

Networks We adopt five pre-trained models, *i.e.,* R101-GeM (Radenović, Tolias, and Chum 2018), R101-AP-GeM (Revaud et al. 2019), R101-SOLAR (Ng et al. 2020), R101-DELG (Cao, Araujo, and Sim 2020; Yang et al. 2021), R101-DOLG (Yang et al. 2021) as teacher models, considering their being open-source and decent performance on the landmark image retrieval task. R refers to ResNet (He et al. 2016). We collect the best-performing versions of these pre-trained models from their open-source repository. For teacher models R101-DELG and R101-DOLG with optional local feature outputs, we only use their global outputs. Shallower networks ResNet-18/34 (R18/34) are utilized as the student model. Note that R18 is the default choice unless otherwise stated.

Influence on Teacher Models' Performance The first concern raised by the utilization of whitening is whether it will compromise the performance of teacher models. Table 1 summarizes the performance of each teacher model with and without whitening. It can be observed that the capacities of teacher models are largely retained after whitening. The limited influence of whitening on the teacher models' performance provides as a preliminary guarantee for multi-teacher distillation using whitened teachers.

Effectiveness of Whitening As shown in Table 2, no matter what fusion strategy is adopted, the whitening group can bring consistent performance gains than the non-whitening group, which indicates that whitening teacher models can boost multi-teacher distillation. Without whitening, the performance shows inconsistency between $\mathcal{R}$Oxf and $\mathcal{R}$Par (*e.g.,* max-mean versus max-rand) and unstable between strategies, which is probably due to the direct fusion of incommensurable similarities. In contrast, whitening makes it less sensitive to the selection of fusion strategies and improves on both datasets.

Empirical Studies on Fusion Strategies Among all the strategies, *max-min* performs the best when whitening is adopted. We attribute it to that *max-min* is the most consistent strategy with the objective of visual retrieval where positive pairs are supposed to be embedded closer together and negative pairs further apart. Taking larger positive similarities (max) and smaller negative similarities (min) will produce a fused similarity matrix with higher quality than that of an individual teacher: similarities between positive pairs are more distinguished from negative pairs, therefore the overall ranks of positive samples can be improved. Additionally, *mean* strategy by taking the average of all similarities produces moderate performance. This demonstrates that taking the average is not always an effective strategy for multi-teacher distillation, though it has been the *de facto* choice in most works (Ba and Caruana 2014; Hinton et al. 2014; You et al. 2017). For the remaining strategies, *i.e., max-mean*, *max-rand*, *rand*, when teachers are not whitened, the performance is at most comparable to whitened counterparts. Based on the results, *max-min* strategy will be adopted in later experiments.

Single-Teacher Distillation vs. Multi-Teacher Distillation Table 3 summarizes the performance and complexity of original teacher models and distillation models using varying teacher combinations. The first group reports the performance of whitened teacher models. As shown in the table, single-teacher distillation could significantly reduce the model complexity during the inference, but their performance is only comparable to the corresponding teacher models. By contrast, multi-teacher distillation methods including double-teacher and triple-teacher distillation not only reduce the model complexity but also achieve further performance gain. We attribute the performance gain to the complementarity of multiple teacher models and the effectiveness of our proposed multi-teacher distillation framework. Besides, the triple-teacher distillation almost surpassing the double-teacher distillation demonstrates the incremental advantage of fusing more teacher models.

It is possible to improve even further by introducing

Method	Teacher models	Params (M)	GFLOPs	$\mathcal{R}$Oxf		$\mathcal{R}$Par		$\mathcal{R}$Oxf+1M		$\mathcal{R}$Par+1M	
				M	H	M	H	M	H	M	H
R101-GeM	-	46.70	124	68.97	45.08	82.15	65.02	55.68	29.24	61.73	35.07
R101-AP-GeM	-	46.70	124	70.37	45.93	81.02	63.29	57.48	32.01	61.26	35.59
R101-SOLAR	-	56.15	139	68.52	43.65	81.82	64.39	56.24	29.65	61.53	35.12
Single-teacher distillation	R101-GeM	11.44	28.62	70.95	46.42	81.68	64.03	55.82	28.77	61.04	33.90
	R101-AP-GeM			71.41	46.46	80.64	62.82	57.61	31.17	59.67	33.57
	R101-SOLAR			69.78	45.12	81.78	63.76	55.87	28.84	60.63	34.04
Double-teacher distillation	R101-GeM, R101-AP-GeM	11.44	28.62	73.90	49.81	84.09	67.91	61.12	33.72	64.44	38.89
	R101-GeM, R101-SOLAR			70.79	46.95	82.48	64.93	56.20	29.41	62.57	36.13
	R101-AP-GeM, R101-SOLAR			**74.71**	50.21	83.59	67.01	60.93	33.01	64.28	39.53
Triple-teacher distillation	R101-GeM, R101-AP-GeM, R101-SOLAR	11.44	28.62	74.67	**50.69**	**84.48**	**68.34**	**61.74**	**34.49**	**64.82**	**39.63**

Table 3: Comparison of original teacher models and distillation models using varying teacher combinations in terms of model complexity and performance. All the distillation variants utilize R18 as the student model. Multi-teacher distillation methods not only reduce the model complexity but also achieve further performance gain.

Method	Params (M)	GFLOPs	$\mathcal{R}$Oxf-M/H	$\mathcal{R}$Par-M/H
EM	149.55	387	71.14/46.67	83.38/66.41
ED	11.44	28.62	68.19/42.29	80.47/61.73
CL	11.44	28.62	65.72/40.25	81.78/63.69
Ours	11.44	28.62	**74.67/50.69**	**84.48/68.34**

Table 4: Comparison to common baseline approaches. Our proposed multi-teacher distillation framework is much more lightweight and effective than optional approaches.

more teacher models. But we can also imagine a performance saturation with endless addition of teachers, due to the student model's limited capacity, less complementarity between teachers, etc. Using more teacher models also means larger training cost. It is a trade-off in practice and we recommend determining the number of teachers based on the expected service time to make the additional training cost more amortized by the improved inference speed.

Comparison to Baseline Approaches In order to further verify the effectiveness of multi-teacher distillation, we compare it to three related baseline approaches, including Ensemble Mean (EM), Embedding Distillation (ED), and Contrastive Learning (CL). EM is a classical ensemble method, which averages the similarity scores produced by teacher models during inference. ED is a feature-based distillation method, which jointly minimizes the Euclidean distances between output embeddings of teacher models and the student model. CL is a representation learning method widely used in recent works (Chen, Xie, and He 2021), which is equivalent to replacing $\mathcal{Q}$ in Eq. 2 with an identity matrix.

As in Table 4, our proposed triple-teacher distillation consistently outperforms the counterparts with a clear margin. Among them, EM fusing knowledge from multiple models performs the best, but it brings heavy computation overhead as every ensembled model must be forwarded to calculate its similarity output. By contrast, our proposed multi-teacher distillation framework only needs to forward a lightweight student once during inference. Additionally, the worse performance of ED than ours demonstrates that relation-based distillation is more suitable for multi-teacher distillation.

Embedding Dimension of Student Model Not only does the computation overhead account for efficient visual retrieval, but also the embedding dimension is a crucial consideration in practice. A smaller embedding dimension of retrieval models is preferred as it can greatly reduce the storage and computation cost. Therefore, we study the influence of the output embedding dimension on the final retrieval performance. We compare to R101-GeM, R101-AP-GeM and R101-SOLAR, which are jointly used as the teacher models in our Whiten-MTD. As depicted in Figure 3, the student model consistently outperforms each teacher model with clear margins. The results highlight the effectiveness of our multi-teacher distillation framework.

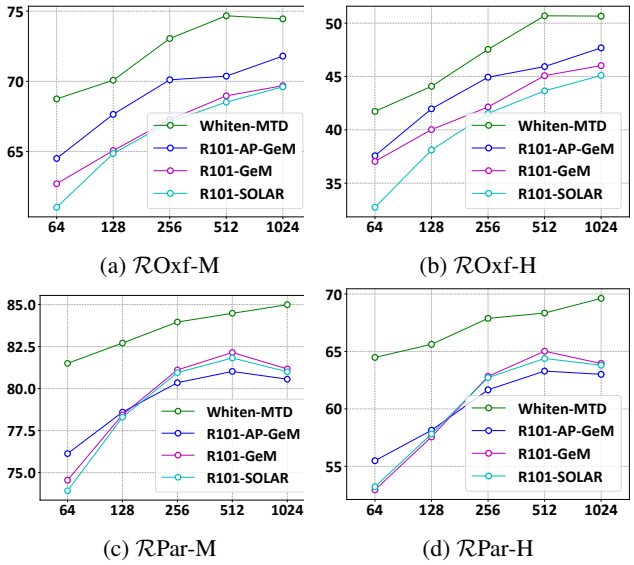
Comparison to State-of-the-Art Table 5 summarizes the performance comparison on $\mathcal{R}$Oxf and $\mathcal{R}$Par, where the state-of-the-art methods are categorized into three groups depending on the type of features used. Note that we only utilize the global features for evaluation, hence our approach belongs to the third group. To make it more comparable to state-of-the-art methods, we choose two best-performing methods R101-DELG and R101-DOLG as teacher models.

Method	Params (M)	GFLOPs	$\mathcal{R}$Oxf		$\mathcal{R}$Par		$\mathcal{R}$Oxf+1M		$\mathcal{R}$Par+1M	
			M	H	M	H	M	H	M	H
(A) Local feature aggregation										
R50-DELF-ASMK+SP (Noh et al.; Radenović et al.)	9.07	53.82	67.80	43.10	76.90	55.40	53.80	31.20	57.30	26.40
R50-DELF-R-ASMK+SP (Teichmann et al.)	9.07	53.82	76.00	52.40	80.20	58.60	64.00	38.10	59.70	29.40
R50-HOW-ASMK (Tolias, Jenicek, and Chum)	8.67	52.61	79.40	56.90	81.60	62.40	65.80	38.90	61.80	33.70
(B) Global features + local features										
R101-GeM↑+DSM (Siméoni, Avrithis, and Chum)	42.50	124	65.30	39.20	77.40	56.20	47.60	23.20	52.80	25.00
R101-DELG (Cao, Araujo, and Sim; Yang et al.)	44.08	125	81.20	64.00	87.20	72.80	69.10	47.50	71.50	48.70
R50-DELG+RRT (Tan, Yuan, and Ordonez)	25.08	66.55	78.10	60.20	86.70	75.10	67.00	44.10	69.80	49.40
(C) Global features										
R101-GeM (Radenović, Tolias, and Chum)	46.70	124	68.43	44.41	82.19	65.05	55.25	30.04	61.77	35.14
R101-AP-GeM (Revaud et al.)	46.70	124	70.00	45.60	80.91	63.00	57.36	31.89	61.09	35.27
R101-SOLAR (Ng et al.)	56.15	139	69.90	47.90	81.60	64.50	53.50	29.90	59.20	33.40
R101-DELG (Cao, Araujo, and Sim; Yang et al.)	43.55	124	76.30	55.60	86.60	72.40	63.70	37.50	70.60	46.90
R101-DOLG (Yang et al.)	46.70	127	81.50	61.10	91.02	80.30	77.43	54.81	83.29	66.69
R50-CVNet-Global+Rerank (Lee et al.)	35.20	64.41	**87.90**	75.60	90.50	80.20	80.70	65.10	82.40	67.30
R101-CVNet-Global+Rerank (Lee et al.)	54.20	123	87.20	**75.90**	91.20	81.10	**81.90**	**67.40**	83.80	69.30
DToP-R50+ViT-B (Song et al.)	-	-	84.40	64.80	92.30	**84.60**	78.90	57.10	**85.40**	**71.20**
R18-Whiten-MTD (ours)	11.44	28.62	81.71	59.73	90.67	80.15	74.05	46.48	77.24	57.90
R34-Whiten-MTD (ours)	21.55	57.71	83.05	64.16	**92.33**	83.20	76.88	52.65	80.57	63.26

Table 5: Comparison to state-of-the-art on $\mathcal{R}$Oxf and $\mathcal{R}$Par datasets. All scores are from their original reports, except R101-DELG which is re-implemented by (Yang et al. 2021) and trained on GLDv2 showing better performance than the original. The best scores are marked in bold. Our models of using only global features are distilled from R101-DOLG and R101-DELG, which demonstrates a good balance of retrieval performance and efficiency.

Method	Params (M)	GFLOPs	mAP@100	mAP
VGG16-CNNL (2017a)	134	15.47	61.04	55.55
VGG16-CNNV (2017a)	134	15.47	25.10	19.09
VGG16-CTE (2013)	134	15.47	-	50.97
VGG16-DML (2017b)	139	15.47	81.27	78.47
R50-VRL (2022)	23.5	4.14	86.00	-
R50-DnS (2022)	27.55	4.13	-	**90.20**
R50-MoCoV3 (2021)	23.5	4.14	87.31	85.47
R50-BarlowTwins (2021)	23.5	4.14	87.22	84.80
R18-Whiten-MTD (ours)	11.2	1.83	88.62	86.82
R34-Whiten-MTD (ours)	21.3	3.68	**88.84**	86.78

Table 6: Comparison to state-of-the-art on SVD. MoCoV3 and BarlowTwins are the teacher models. Whiten-MTD with lightweight backbones achieves comparable performance.



(a) $\mathcal{R}$Oxf-M (b) $\mathcal{R}$Oxf-H

(c) $\mathcal{R}$Par-M (d) $\mathcal{R}$Par-H

Figure 3: Performance of varying output dimension on $\mathcal{R}$Oxf and $\mathcal{R}$Par datasets. Our proposed WhitenMTD consistently outperforms them with clear margins.

Compared to existing methods based on the deep R50/R101 as the backbone, our Whiten-MTD achieves comparable performance even if R18 is utilized as the student model. Using a larger R34 as the student model brings further performance boost, which achieves the best result on $\mathcal{R}$Par-M and competitive results otherwise. Even compared against reranking methods, our model still obtains competitive results with fewer parameters and GFLOPs. It is worth pointing out that models in the first group have fewer parameters than ours due to the last block being removed from R50, but their performance is much worse. What's more, when distilling from the two best-performing R101-DELG and R101-DOLG, the students are slightly worse than the teachers (see Table 1) but instead more efficient. The results demonstrate that our Whiten-MTD accomplishes a good balance of retrieval performance and efficiency.

Near-Duplicate Video Retrieval

Experimental Setup We utilize Short Video Dataset (SVD) (Jiang et al. 2019), considering it is the latest and the largest benchmark dataset for near-duplicate video retrieval. We collect two self-supervised pre-trained models as teachers, *i.e.,* R50-MoCoV3 (Chen, Xie, and He 2021), R50-BarlowTwins (Zbontar et al. 2021). For the student model, we also use R18 and R34. To make a fair comparison with state-of-the-art works (He et al. 2022; Kordopatis-Zilos et al. 2022), we report both mAP@100 and mAP.

Results The results on the SVD dataset are shown in Table 6. With a more lightweight backbone, our model achieves comparable performance with mAP@100 of 88.84 and mAP of 86.82. State-of-the-art method (Kordopatis-Zilos et al. 2022) relies on computationally expensive spatial-temporal feature extraction and reranking to achieve a better result while we only perform a simple forward of R18 or R34. The results demonstrate the effectiveness of our multi-teacher distillation framework for video retrieval.

Conclusion

In this paper we contribute a multi-teacher distillation method for efficient visual retrieval. We propose a simple and effective multi-teacher distillation framework Whiten-MTD. We also investigate five heuristic fusion strategies and make a detailed analysis of their adaptations. For effective multi-teacher distillation whitening is needed before applying any fusion strategies to eliminate the similarity distribution discrepancy of teacher models. Extensive experiments demonstrate the student models obtained using our method are computationally efficient, with comparable performance to state-of-the-art methods based on heavy networks on both instance image retrieval and video retrieval.

Acknowledgments

This work is sponsored by the National Key Research, Development Program of China under No. 2022YFB3102100, Pioneer and Leading Goose R&D Program of Zhejiang (No. 2023C01212), Young Elite Scientists Sponsorship Program by CAST (No. 2022QNRC001), National Natural Science Foundation of China (No. 61976188, No. 62372402), CCF-AFSG Research Fund, R&D Program of DCI Technology Application Joint Laboratory.

References

Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; and Sivic, J. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *CVPR*, 5297–5307.

Ba, J.; and Caruana, R. 2014. Do deep nets really need to be deep? In *NeurIPS*, 2654–2662.

Beyer, L.; Zhai, X.; Royer, A.; Markeeva, L.; Anil, R.; and Kolesnikov, A. 2022. Knowledge distillation: A good teacher is patient and consistent. In *CVPR*, 10925–10934.

Buciluǎ, C.; Caruana, R.; and Niculescu-Mizil, A. 2006. Model compression. In *SIGKDD*, 535–541.

Cai, T. T.; Fan, J.; and Jiang, T. 2013. Distributions of angles in random packing on spheres. *J. Mach. Learn. Res.*, 14(1): 1837–1864.

Cao, B.; Araujo, A.; and Sim, J. 2020. Unifying deep local and global features for image search. In *ECCV*, 726–743.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *ICCV*, 9650–9660.

Chen, W.; Liu, Y.; Wang, W.; Bakker, E. M.; Georgiou, T.; Fieguth, P.; Liu, L.; and Lew, M. S. 2023. Deep learning for instance retrieval: A survey. *TPAMI*, 45(6): 7270–7292.

Chen, X.; Xie, S.; and He, K. 2021. An empirical study of training self-supervised vision transformers. In *ICCV*, 9640–9649.

Dong, J.; Li, X.; Xu, C.; Yang, G.; and Wang, X. 2018. Feature re-learning with data augmentation for content-based video recommendation. In *MM*, 2058–2062.

Dong, J.; Ma, Z.; Mao, X.; Yang, X.; He, Y.; Hong, R.; and Ji, S. 2021. Fine-grained fashion similarity prediction by attribute-specific embedding learning. *TIP*, 30: 8410–8425.

Dong, J.; Zhang, M.; Zhang, Z.; Chen, X.; Liu, D.; Qu, X.; Wang, X.; and Liu, B. 2023. Dual Learning with Dynamic Knowledge Distillation for Partially Relevant Video Retrieval. In *ICCV*, 11302–11312.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An image is worth 16x16 words: transformers for image recognition at scale. In *ICLR*.

Duggal, R.; Zhou, H.; Yang, S.; Xiong, Y.; Xia, W.; Tu, Z.; and Soatto, S. 2021. Compatibility-aware heterogeneous visual search. In *CVPR*, 10723–10732.

Dusmanu, M.; Rocco, I.; Pajdla, T.; Pollefeys, M.; Sivic, J.; Torii, A.; and Sattler, T. 2019. D2-net: A trainable cnn for joint detection and description of local features. In *CVPR*, 8092–8101.

Ermolov, A.; Siarohin, A.; Sangineto, E.; and Sebe, N. 2021. Whitening for self-supervised representation learning. In *ICML*, 3015–3024.

Fang, Z.; Wang, J.; Wang, L.; Zhang, L.; Yang, Y.; and Liu, Z. 2021. SEED: Self-supervised distillation for visual representation. In *ICLR*.

Fischler, M. A.; and Bolles, R. C. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6): 381–395.

Gordo, A.; Almazán, J.; Revaud, J.; and Larlus, D. 2016. Deep image retrieval: Learning global representations for image search. In *ECCV*, 241–257.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. In *NeurIPS*, volume 33, 21271–21284.

Guo, S.; Alvarez, J. M.; and Salzmann, M. 2021. Distilling image classifiers in object detectors. In *NeurIPS*, 1036–1047.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

He, X.; Pan, Y.; Tang, M.; and Lv, Y. 2021. Self-supervised video retrieval transformer network. *arXiv preprint arXiv:2104.07993*.

He, X.; Pan, Y.; Tang, M.; Lv, Y.; and Peng, Y. 2022. Learn from unlabeled videos for near-duplicate video retrieval. In *SIGIR*, 1002–1011.

Hinton, G.; Vinyals, O.; Dean, J.; et al. 2014. Distilling the knowledge in a neural network. In *NeurIPS 2014 Workshop*.

Jiang, Q.-Y.; He, Y.; Li, G.; Lin, J.; Li, L.; and Li, W.-J. 2019. SVD: A large-scale short video dataset for near-duplicate video retrieval. In *ICCV*, 5281–5289.

Kang, Z.; Zhang, P.; Zhang, X.; Sun, J.; and Zheng, N. 2021. Instance-conditional knowledge distillation for object detection. In *NeurIPS*, 16468–16480.

Kim, J.; Yu, Y.; Kim, H.; and Kim, G. 2021. Dual compositional learning in interactive image retrieval. In *AAAI*, 1771–1779.

Kordopatis-Zilos, G.; Papadopoulos, S.; Patras, I.; and Kompatsiaris, Y. 2017a. Near-duplicate video retrieval by aggregating intermediate cnn layers. In *MMM*, 251–263.

Kordopatis-Zilos, G.; Papadopoulos, S.; Patras, I.; and Kompatsiaris, Y. 2017b. Near-duplicate video retrieval with deep metric learning. In *ICCV Workshops*, 347–356.

Kordopatis-Zilos, G.; Tzelepis, C.; Papadopoulos, S.; Kompatsiaris, I.; and Patras, I. 2022. DnS: Distill-and-select for efficient and accurate video indexing and retrieval. *IJCV*, 130(10): 2385–2407.

Lee, S.; Seong, H.; Lee, S.; and Kim, E. 2022. Correlation verification for image retrieval. In *CVPR*, 5374–5384.

Liu, B.; Zheng, T.; Zheng, P.; Liu, D.; Qu, X.; Gao, J.; Dong, J.; and Wang, X. 2023. Lite-MKD: A Multi-modal Knowledge Distillation Framework for Lightweight Few-shot Action Recognition. In *MM*, 7283–7294.

Liu, J.; Huang, Z.; Cai, H.; Shen, H. T.; Ngo, C. W.; and Wang, W. 2013. Near-duplicate video retrieval: Current research and future trends. *CSUR*, 45(4): 1–23.

Luo, Z.; Shen, T.; Zhou, L.; Zhang, J.; Yao, Y.; Li, S.; Fang, T.; and Quan, L. 2019. Contextdesc: Local descriptor augmentation with cross-modality context. In *CVPR*, 2527–2536.

Ng, T.; Balntas, V.; Tian, Y.; and Mikolajczyk, K. 2020. SOLAR: second-order loss and attention for image retrieval. In *ECCV*, 253–270.

Noh, H.; Araujo, A.; Sim, J.; Weyand, T.; and Han, B. 2017. Large-scale image retrieval with attentive deep local features. In *ICCV*, 3456–3465.

Noroozi, M.; Vinjimoor, A.; Favaro, P.; and Pirsiavash, H. 2018. Boosting self-supervised learning via knowledge transfer. In *CVPR*, 9359–9367.

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational knowledge distillation. In *CVPR*, 3967–3976.

Passalis, N.; and Tefas, A. 2018. Learning deep representations with probabilistic knowledge transfer. In *ECCV*, 268–284.

Passalis, N.; Tzelepi, M.; and Tefas, A. 2020. Heterogeneous knowledge distillation using information flow modeling. In *CVPR*, 2339–2348.

Philbin, J.; Chum, O.; Isard, M.; Sivic, J.; and Zisserman, A. 2007. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 1–8.

Philbin, J.; Chum, O.; Isard, M.; Sivic, J.; and Zisserman, A. 2008. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, 1–8.

Radenović, F.; Iscen, A.; Tolias, G.; Avrithis, Y.; and Chum, O. 2018. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *CVPR*, 5706–5715.

Radenović, F.; Tolias, G.; and Chum, O. 2018. Fine-tuning CNN image retrieval with no human annotation. *TPAMI*, 41(7): 1655–1668.

Revaud, J.; Almazán, J.; Rezende, R. S.; and Souza, C. R. d. 2019. Learning with average precision: Training image retrieval with a listwise loss. In *ICCV*, 5107–5116.

Revaud, J.; Douze, M.; Schmid, C.; and Jégou, H. 2013. Event retrieval in large video collections with circulant temporal encoding. In *CVPR*, 2459–2466.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2015. Fitnets: Hints for thin deep nets. In *ICLR*.

Shao, J.; Wen, X.; Zhao, B.; and Xue, X. 2021. Temporal context aggregation for video retrieval with contrastive learning. In *WACV*, 3268–3278.

Shen, L.; Hong, R.; and Hao, Y. 2020. Advance on large scale near-duplicate video retrieval. *FCS*, 14(5): 1–24.

Shen, Z.; He, Z.; and Xue, X. 2019. Meal: Multi-model ensemble via adversarial learning. In *AAAI*, 4886–4893.

Siméoni, O.; Avrithis, Y.; and Chum, O. 2019. Local features and visual words emerge in activations. In *CVPR*, 11651–11660.

Song, C. H.; Yoon, J.; Choi, S.; and Avrithis, Y. 2023. Boosting vision transformers for image retrieval. In *WACV*, 107–117.

Tan, F.; Yuan, J.; and Ordonez, V. 2021. Instance-level image retrieval using reranking transformers. In *ICCV*, 12105–12115.

Teichmann, M.; Araujo, A.; Zhu, M.; and Sim, J. 2019. Detect-to-retrieve: Efficient regional aggregation for image search. In *CVPR*, 5109–5118.

Tian, Y.; Krishnan, D.; and Isola, P. 2019. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*.

Tolias, G.; Jenicek, T.; and Chum, O. 2020. Learning and aggregating deep local descriptors for instance-level recognition. In *ECCV*, 460–477.

Tolias, G.; Sicre, R.; and Jégou, H. 2015. Particular object retrieval with integral max-pooling of CNN activations. *arXiv preprint arXiv:1511.05879*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*, 5998–6008.

Weng, X.; Huang, L.; Zhao, L.; Anwer, R.; Khan, S. H.; and Shahbaz Khan, F. 2022. An investigation into whitening loss for self-supervised learning. In *NeurIPS*, 29748–29760.

Weyand, T.; Araujo, A.; Cao, B.; and Sim, J. 2020. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *CVPR*, 2575–2584.

Wu, H.; Wang, M.; Zhou, W.; Li, H.; and Tian, Q. 2022. Contextual similarity distillation for asymmetric image retrieval. In *CVPR*, 9489–9498.

Yang, M.; He, D.; Fan, M.; Shi, B.; Xue, X.; Li, F.; Ding, E.; and Huang, J. 2021. DOLG: Single-stage image retrieval with deep orthogonal fusion of local and global features. In *ICCV*, 11772–11781.

Yim, J.; Joo, D.; Bae, J.; and Kim, J. 2017. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *CVPR*, 4133–4141.

You, S.; Xu, C.; Xu, C.; and Tao, D. 2017. Learning from multiple teacher networks. In *SIGKDD*, 1285–1294.

Yuan, F.; Shou, L.; Pei, J.; Lin, W.; Gong, M.; Fu, Y.; and Jiang, D. 2021. Reinforced multi-teacher selection for knowledge distillation. In *AAAI*, 14284–14291.

Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; and Deny, S. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, 12310–12320.

Zhixing, D.; Zhang, R.; Chang, M.; Liu, S.; Chen, T.; Chen, Y.; et al. 2021. Distilling object detectors with feature richness. In *NeurIPS*, 5213–5224.