Coarse-to-Fine: Learning Compact Discriminative Representation for Single-Stage Image Retrieval

Yunquan Zhu*, Xinkai Gao*, Bo Ke, Ruizhi Qiao, Xing Sun
YouTu Lab, Tencent, China
{yunquanzhu, kayxgao, boke, ruizhiqiao, winfredsun}@tencent.com

Abstract

Image retrieval targets to find images from a database that are visually similar to the query image. Two-stage methods following retrieve-and-rerank paradigm have achieved excellent performance, but their separate local and global modules are inefficient to real-world applications. To better trade-off retrieval efficiency and accuracy, some approaches fuse global and local feature into a joint representation to perform single-stage image retrieval. However, they are still challenging due to various situations to tackle, e.g.*, background, occlusion and viewpoint. In this work, we design a* **C***oarse-to-***F***ine framework to learn* **C***ompact* **D***iscriminative representation (CFCD) for end-to-end single-stage image retrieval-requiring only image-level labels. Specifically, we first design a novel adaptive softmax-based loss which dynamically tunes its scale and margin within each mini-batch and increases them progressively to strengthen supervision during training and intra-class compactness. Furthermore, we propose a mechanism which attentively selects prominent local descriptors and infuse fine-grained semantic relations into the global representation by a hard negative sampling strategy to optimize inter-class distinctiveness at a global scale. Extensive experimental results have demonstrated the effectiveness of our method, which achieves state-of-the-art single-stage image retrieval performance on benchmarks such as Revisited Oxford and Revisited Paris. Code is available at https://github.com/bassyess/CFCD.*

1. Introduction

Image retrieval is a fundamental task in computer vision, which aims to efficiently retrieve images similar to a given query from a large-scale database. With the development of deep learning, image retrieval has made great progress [24, 33, 41, 9, 5]. The state-of-the-art methods generally work in a two-stage paradigm [4, 21], where they first obtain coarse candidates via global features, and then re-rank them with local features to achieve better performance. However, two-stage methods are required to rank images twice and use the expensive RANSAC [12] or AMSK [35] for geometric verification, leading to high memory usage and increased latency.

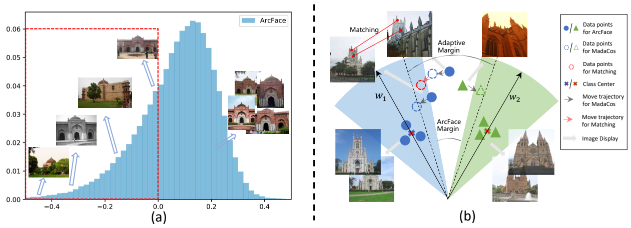


Figure 1. (a) Distribution of $[cos(\theta_{y_i}+m)-argmax(cos\theta_{j\neq y_i})]$ after training with ArcFace, where y_i is target label. In the red box, due to large variations in background, occlusion and viewpoint, these images in Google Landmarks Dataset V2 are far away from their class centers and misclassified. (b) Geometrical interpretation of our methods from the feature perspective. By designing an adaptive margin penalty strategy, we can introduce appropriate supervision intensity for different batch during training. As for the outliers with partial match, we design a mechanism to select prominent local descriptors and minimize their pairwise distances, which makes the unified representation more discriminative.

To alleviate the efficiency issues, many studies [13, 23, 20, 11] recently attempt to explore a unified single-stage image retrieval solution. They design complicated attention modules to fuse global and local features, and adopt the ArcFace [8] loss to train the model in an end-to-end fashion. They have shown excellent performance on single-stage image retrieval benchmarks. In spite of their successes, extracting multi-scale local features is still an extremely expensive process. More importantly, these studies do not consider the challenges of large-scale landmark dataset from the perspective of data distribution, which have large variations in background, occlusion and viewpoint.

Fig1(a) displays the cosine logits distribution of landmark samples after convergence, where more than 20% of the samples are far away from their class centers as their

*Equal contribution.

target cosine logits are smaller than their non-target cosine logits. One can observe the various conditions in these samples such as background, occlusion, viewpoint, *etc*. Moreover, true positives lingering at the classification boundary receive weaker supervision due to the fixed margin penalty. Therefore, we propose an adaptive margin penalty strategy that tunes hyper-parameters to progressively strengthen supervision for intra-class compactness. Besides, inspired by geometric verification, in order to retrieve target images with partial match, we design a mechanism to select prominent local descriptors and minimize their pairwise distances for learning inter-class distinctiveness more effectively, is shown in Fig1(b).

We propose a **C**oarse-to-**F**ine framework to learn **C**ompact and **D**iscriminative representation (CFCD) for single-stage image retrieval. Specifically, we first propose a novel adaptive loss which uses the median of cosine logits in a batch to dynamically tune the scale and margin of the loss function, namely MadaCos. MadaCos increases its scale and margin progressively to strengthen supervision during training, consequently increasing the learned intra-class compactness. We also design the local descriptors matching constraints and hard negative sampling strategy to construct triplets, and introduce the triplet loss[3] to leverage fine-grained semantic relations, which embed the global feature with more information of inter-class distinctiveness. We jointly train the model as a whole with MadaCos and triplet losses to produce the final compact and discriminative representation and improve the overall performance. This framework consists of two training phases: global feature learning with MadaCos and later added local feature matching with triplet loss. During the testing stage, global features are extracted from the the end-to-end framework and ranked once without additional computation overhead. Our main contributions are summarized as follows:

- We propose a coarse-to-fine framework to learn compact and discriminative representation for single-stage image retrieval without additional re-ranking computation overhead, which is more efficient.
- To enhance intra-class compactness, we design an adaptive softmax loss named MadaCos, which uses the median of cosine logits within each mini-batch to tune its hyperparameters to strengthen supervision.
- To enhance inter-class distinctiveness, we select prominent local descriptors and design an image-level hard negative sampling strategy to leverage fine-grained semantic relations.
- Through systematic experiments, the proposed method achieves stage-of-the-art single-stage image retrieval performance on benchmarks: $\mathcal{R}$Oxf (+1M), $\mathcal{R}$Par (+1M).

2. Related Work

2.1. Image Retrieval

In early researches, global features are developed by aggregating hand-crafted local features through Fisher vector [19], VLAD [18] or ASMK [35]. Afterward, spatial verification performs local features matching with RANSAC [12] to re-rank preliminary retrieval results, which effectively improves the overall performance. Recently, handcrafted features have been replaced by global and local features extracted from deep learning networks. In local features based image retrieval, [2, 44, 16, 10, 39] have made remarkable progress by leveraging discriminative geometry information. From the global aspect, high-level semantic features are obtained simply by performing differentiable pooling operations such as sum-pooling(SPoC)[1], regional-max-pooling(R-MAC)[14] and generalized mean-pooling(GeM)[30] on the feature maps of CNNs. The state-of-the-art approaches leverage local and global features to explore two types of algorithm pipelines. In the two-stage paradigm, the typical method DELG [4] incorporated with DELF's [26] attention module trains local and global features in an end-to-end manner. They first search by global features, then re-rank the top database images using local feature matching. To alleviate the problems of high memory usage and latency of two-stage methods, the single-stage methods such as Token [41] and DOLG[42] fuse local features or both local and global features into a compact representation, and only rank images once. However, the two categories either suffer from inefficient local branches or weak supervisions of local features. Our work is essentially different from them. We propose a coarse-to-fine framework to quickly train coarse global and local descriptors, and then select matching local descriptors to refine the global features to integrate local fine-grained information. In other words, we introduce the idea of geometric validation to supervise local features in an end-to-end training scheme, which replaces the complicated local branches. Moreover, the inference of our framework is performed without additional re-ranking computation overhead.

2.2. Margin-based Softmax Loss

For learning deep image representations, previous works propose pair-based losses such as triplet [3], angular [38] and listwise [31] losses to train CNN models. Recent works propose margin-based losses such as SphereFace[22], CosFace[37], ArcFace[8],and Sub-center ArcFace[7], which maximize angular margin to the target logit and thus lead to faster convergence and better performance. Among them, the state-of-the-art studies[40, 42] directly adopt ArcFace loss to train the whole model. However, the training process of ArcFace loss is usually tricky and unstable, so one has to repeat training with multiple set-

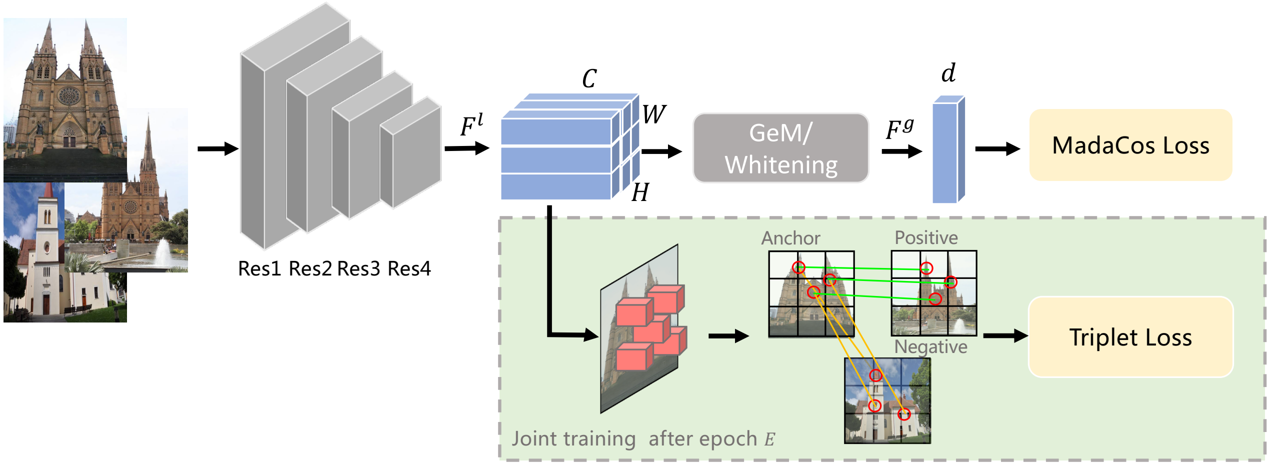


Figure 2. Illustration of the proposed coarse-to-fine framework to learn compact discriminative representation (CFCD) and its training objectives. The components highlighted in green are introduced after training for E epochs. We use MadaCos alone to train global features for the first E epochs, then select the attention regions from the attention maps as constraints to construct triplets for local descriptors matching, and finally train the model with both MadaCos and triplet losses.

tings to achieve optimal performance. Adacos[43] attempt to leverage adaptive or learnable scale and margin parameters, but they pay less attention to the softmax function curve and data characteristics, *e.g.* large variations in background, occlusion and viewpoint. Therefore, we propose a MadaCos loss which automatically tunes its hyperparameters to perform more accurate single-stage image retrieval.

3. Methods

3.1. Overview

Our CFCD framework is depicted in Fig.2. Given an image, we obtain the original deep local descriptors $F^l \in \mathbb{R}^{d_c \times d_w \times d_h}$ via a CNN backbone, where d_c, d_w and d_h are the dimensions of channels, width and height of the feature map, respectively. We then use GeM pooling and a whitening FC layer to extract the global representation $F^g \in \mathbb{R}^{d_g}$ of a dimension d_g. We propose MadaCos, an adaptive softmax-based loss to learn the global representation. During the first E training epochs, MadaCos is used alone to make the network aware of prominent local regions in F^l. The prominent regions are selected from the attention maps as matching constraints, which prompt us to design a hard negative sampling strategy to construct triplets for local descriptor matching. After E epochs, a triplet loss is combined with MadaCos to jointly train the network so that the global features are infused with geometry information about discriminative local regions.

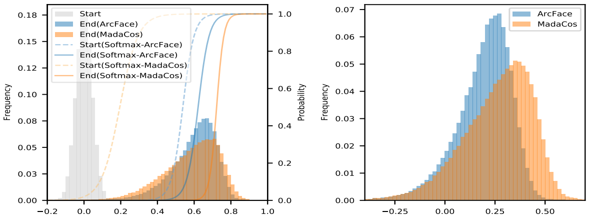
3.2. Global Feature Learning with MadaCos Loss

Let x_i be the i-th image of the current mini-batch with size N, and y_i the corresponding label. Margin-based softmax losses [22, 37, 8] apply ℓ_2 normalization to the classifier weight and the embedding feature. They use three kinds of margin penalty, *i.e.*, multiplicative angular margin m_1, additive angular margin m_2, and additive cosine margin m_3, respectively. We denote θ_{y_i} as the angle between the target weight and the feature, and $cos\theta_{y_i}$ is its cosine logit. Then the margin-based softmax losses can be combined into a unified framework:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{s(cos(m_1\theta_{y_i}+m_2)-m_3)}}{e^{s(cos(m_1\theta_{y_i}+m_2)-m_3)} + B_i}, \quad (1)$$

where s is a scale factor, and $B_i = \sum_{j=1,j\neq y_i}^{n} e^{s\, cos\theta_j}$ is the summation of the cosine logits of non-target classes.

Previous works[40, 42] adopt ArcFace loss with additive angular margin m_2 to train the global descriptors, which achieves better performance than other margin-based softmax functions. According to Eq.1, we show the distribution of $cos\theta_{y_i}$ between embedding feature and the corresponding target center as well as the softmax function curves at the start and end of training in Fig.3(a). As the training converges, θ_{y_i} gradually shrinks so its $cos\theta_{y_i}$ distribution and softmax function gradually shift to the right side. ArcFace loss makes the distribution of class centers scattered and the distribution of $cos\theta_{y_i}$ more concentrated, which intuitively indicates that it enhances the intra-class compactness. Nevertheless, most samples distribute on the right side of the softmax function. This leads to suboptimal performance due to weak supervisions. We therefore propose a novel adaptive loss, namely MadaCos, which automatically tunes appropriate parameters within each mini-batch by imposing strict constrains on the probability of the median of cosine logits to progressively strengthen supervision through-



(a) Cos(θ) between the Feature and Target Center

(b) Distribution of $[Cos(\theta_{y_i})\text{-}Max(Cos(\theta_{j \neq y_i}))]$, y_i is target label

Figure 3. (a) Target cosine logit distributions and softmax function at the start and end of training. (b) Distributions of difference between the target cosine logit and its corresponding maximum of non-target cosine logit after training convergence.

out the training process.

To simplify the derivation of subsequent formulas, we adopt the additive cosine margin m_3 instead of additive angular margin m_2 to impose penalties, and the subscript in m_3 is omitted in the following for simplicity. Formally, we have:

$$P_{i,y_i} = \frac{e^{s(cos\theta_{y_i} - m)}}{e^{s(cos\theta_{y_i} - m)} + B_i}, \tag{2}$$

where P_{i,y_i} represents its probability of assigning x_i to class y_i. And we also introduce a modulating indicator variable $cos_m\theta$, which is the median of target cosine logits in the current mini-batch, and its corresponding label is y_k. $cos_m\theta$ reflects the convergence degree of the under-training network in the the mini-batch. Since the cross entropy loss is $-log(P)$, if we can control the probability $P(cos_m\theta)$ of the median target cosine logit, we potentially control the overall supervision intensity. Therefore, we propose to dynamically compute the appropriate scale s and margin m within each mini-batch so that $P(cos_m\theta)$ reaches an anchor point ρ, which keeps most samples distributing on the left side of the softmax function. Accordingly, the network can impose stronger constraints to progressively enhance supervision with suitable ρ even if the angle θ_{i,y_i} shrinks. Based on this observation, we set the $P(cos_m\theta) = \rho$ to compute scale s and margin m. Here ρ is set to 0.02 in our experiments,

$$\frac{e^{s(cos_m\theta - m)}}{e^{s(cos_m\theta - m)} + \tilde{B}} = \rho, \tag{3}$$

Eq.3 ensures that the margin m and scale s gradually increase to avoid model divergence due to too large margin and scale at the early phase of training. However, if s is too small ($e.g.$, $s = 10$), P_{i,y_i} will be very low when θ_{y_i} is close to 0, which means that the loss function may still penalize correctly classified samples. Therefore, we force P_{i,y_i} to be close to 1 when the angle θ_{y_i} is 0.

$$\frac{e^{s(1-m)}}{e^{s(1-m)} + \tilde{B}} = 1 - \epsilon, \tag{4}$$

where $\tilde{B} = \sum_{j=1, j \neq y_k}^{n} e^{s\,cos\theta_j}$. We set $\epsilon = e^{-7}$ in our experiments. Combining Eq.3 and Eq.4, we can derive s and m within each mini-batch to update the MadaCos loss to train the global descriptors progressively. The entire process is summarized in Algorithm 1.

Algorithm 1 MadaCos

Input: The image x_i of i-th sample with label y_i in the mini-batch with size N, the target cosine logit $cos\theta_{y_i}$
Parameter: scale s and margin m

```
$cos_m\theta = Median\{cos\theta_{y_0}, cos\theta_{y_1}, \dots, cos\theta_{y_{N-1}}\}$;
Substitute $cos_m\theta$ into Eq.3 and Eq.4 to compute scale $s$;
    $s = \frac{log((1-\epsilon)(1-\rho)/(\rho\epsilon))}{1 - cos_m\theta}$
Assuming that the corresponding label of $cos_m\theta$ is $y_k$, substitute $s$ to compute $\tilde{B}_k = \sum_{j \neq y_k}^{n} e^{s\,cos\theta_j}$;
Compute the margin $m$ with scale $s$ and $\tilde{B}_k$ by Eq.4;
    $m = cos_m\theta - \frac{log(\rho\tilde{B}_k/(1-\rho))}{s}$
Update the loss $\mathcal{L}_{mda}$ with $m$ and $s$;
    $\mathcal{L}_{mda} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{s(cos\theta_{y_i} - m)}}{e^{s(cos\theta_{y_i} - m)} + B_i}$
```

Output: loss $\mathcal{L}_{mda}$.

As shown in Fig.3(a), compared with ArcFace, the $cos\theta_{y_i}$ distributions of MadaCos more scattered and most samples are distributed on the left side of the softmax curve. We also plot the distribution of $[cos\theta_{y_i} - argmax(cos\theta_{j \neq y_i})]$ in Fig.3(b), which intuitively illustrates that MadaCos loss has a larger tolerance margin.

3.3. Local Feature Matching with Triplet Loss

Let $f \in \mathbb{R}^{d_c}$ be a local descriptor from local descriptors F^l, and then F^l can be seen as set of $Z = d_w \times d_h$ feature vectors denoted by $\mathcal{F} = \{f_i \in \mathbb{R}^{d_c} : i \in 1 \dots Z\}$. We select prominent local descriptors to minimize their pairwise distance. Obviously, if we select matching local descriptors only relying on nearest neighbor-based constraints, the local descriptors may focus on the non-significant parts such as backgrounds and distractions. Therefore, we introduce attention maps to guide the networks to focus on more semantically salient regions and discard the redundant information. Here we define the function $\eta(f, \mathcal{F}) = argmin_{\forall f_i \in \mathcal{F}} ||f - f_i||_2$ which returns the nearest neighbor of f from set $\mathcal{F}$. And let $\psi(\tau, \mathcal{F}) = \{argmax^{\tau}_{\forall f_i \in \mathcal{F}} ||f_i||\}$ be the attention selective function that returns the top τ percent local descriptors by l_1 norm, where τ is a controlling factor. Now, given a positive pair of images x_a, x_p, the corresponding local descriptors are F^l_a and F^l_p. For any eligible local descriptors $v \in \mathcal{F}$, in order to select matching descriptors (v_a, v_p) of positive pair, we set the following constraints: a) (v_a, v_p) must be reciprocal nearest neighbors, b) they need to be in the attention regions specified by $\psi(\tau, \mathcal{F})$. Let $\mathcal{M}$ be the set of eligible pairs, the conditions above can be formulated as:

$$(v_a, v_p) \in \mathcal{M} \iff \begin{cases} v_a = \eta(v_p, \psi(\tau, F_a^l)) \\ v_p = \eta(v_a, \psi(\tau, F_p^l)) \end{cases} \tag{5}$$

Once we construct all eligible local descriptors $\mathcal{M}$ of positive pairs, we introduce the triplet loss to leverage rich local relations between matches. For Q negative images $x_{n_j}, j = 1, \ldots, Q$, we only rely on the nearest neighbor criterion to select those closest to the anchor image matches. We denote $v_{n_j} = \eta(v_a, F_{n_j}^l)$ as the negative local descriptors extracted from x_{n_j}. The triplet loss can be written as:

$$\mathcal{L}_{trip} = \sum_{(v_a, v_p) \in \mathcal{M}} \sum_{j=1}^{Q} \{||v_a - v_p||_2^2 - ||v_a - v_{n_j}||_2^2 + \mu\}^+, \tag{6}$$

where $\mu = 0.1$ and $Q = 6$ in our experiments.

However, triplets constructed by random sampling do not provide sufficiently strong supervision for this task. To not only keep the accurate prediction of the normal and easy samples, but also make the model concentrate on learning from hard samples, we design a custom global sampling strategy with hard negative samples. The approach is shown in Fig.4, and the detailed sampling strategy is provided in the supplementary material. With the help of the model trained at a sufficient stage, we can use its prediction to ensure that each batch of triplets contain appropriate positives which share common patch-level matches between them while focusing on hard negatives. The sampling strategy is crucial to select hard negative triplets and therefore contributes to improving the overall performance. Unlike previous work [21] which selects negatives in the order of global descriptor matching scores with MoCo-like [15] momentum queue, we select negatives from the whole dataset at each epoch without additional computation.

Finally, the total loss of our backbone network $\mathcal{L}_{tot}$ is the weighted sum of the classification loss $\mathcal{L}_{mda}$ and triplet loss $\mathcal{L}_{trip}$:

$$\mathcal{L}_{tot} = \mathcal{L}_{mda} + \lambda \mathcal{L}_{trip}, \tag{7}$$

where λ is set to 0.05 during training.

4. Experiments

4.1. Implementation Details

Datasets and Evaluation Metric The clean version of Google landmarks dataset V2 (GLDv2-clean) [40] contains 1,580,470 images and 81,313 classes. We randomly divide 80% of the data for training and the rest 20% for validation following previous works[4, 42]. To evaluate our model, we primarily use $\mathcal{R}$Oxford5k [27, 29] and $\mathcal{R}$Paris6k [28, 29] datasets, denoted as $\mathcal{R}$Oxf and $\mathcal{R}$Par. Both datasets comprise 70 queries and include 4993 and 6322 database images, respectively. In addition, an $\mathcal{R}$1M dataset [29] which contains one million distractor images is used for measuring the large-scale retrieval performance. For a fair comparison, mean average precision (mAP) is used as our evaluation metric on both datasets with the medium and hard difficulty protocols.

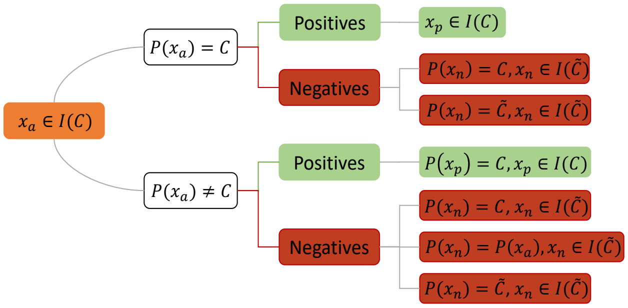

$x_a \in I(C)$
$P(x_a) = C$
Positives
$x_p \in I(C)$
Negatives
$P(x_n) = C, x_n \in I(\tilde{C})$
$P(x_n) = \tilde{C}, x_n \in I(\tilde{C})$
$P(x_a) \neq C$
Positives
$P(x_p) = C, x_p \in I(C)$
Negatives
$P(x_n) = C, x_n \in I(\tilde{C})$
$P(x_n) = P(x_a), x_n \in I(\tilde{C})$
$P(x_n) = \tilde{C}, x_n \in I(\tilde{C})$

Figure 4. Hard negative sampling strategy. Given an anchor image x_a with category C and its prediction $P(x_a)$. $I(C)$ and $I(\tilde{C})$ are the sets of images with category C and non-category $\tilde{C}$, respectively. If $P(x_a) = C$, we randomly sample positives from set $I(C)$ and evenly select negatives from the two predictions of images $x_n \in I(\tilde{C})$. However, if $P(x_a) \neq C$, the anchor image itself is a hard or noise sample, we require that the positives must be select from set $I(C)$ and it's prediction $P(x_p) = C$. The negatives are also evenly selected from the three predictions of images $x_n \in I(\tilde{C})$.

Training Details ResNet50 and ResNet101[17] are mainly used for experiments. Models in this paper are initialized from Imagenet[6] pre-trained weights. The images first undergo augmentations include random cropping and aspect ratio distortion, then are resized to 512×512 following previous works[4, 42]. The models are trained on 8 V100 GPUs for T epochs with the batch size of 128. The initial learning rate of is 0.01. We use SGD optimizer with momentum of 0.9, and set weight decay factor to 0.0001. We also adopt the cosine learning rate decay strategy in the first E epochs to train with MadaCos, and after the E-th epoch, we reset the learning rate to 0.005 to continue training the model with both MadaCos and triplet losses for the remaining $T - E$ epochs. For GeM pooling, we fix the parameter $\boldsymbol{p}$ as 3.0. As for global feature extraction, we also produce multi-scale representations. And ℓ_2 normalization is applied for each scale independently then they are average-pooled, followed by another ℓ_2 normalization step. We use two kinds of experimental settings for fair comparisons. For comparing with DOLG and FIRe[39], we set T to 100, E to 50, $\boldsymbol{d_g}$ to 512 and use 5 scales, $\{\frac{1}{2\sqrt{2}}, \frac{1}{2}, \frac{1}{\sqrt{2}}, 1, \sqrt{2}\}$. For comparing with other methods, we set T to 25, E to 20, $\boldsymbol{d_g}$ to 2048 and use 3 scales, $\{\frac{1}{\sqrt{2}}, 1, \sqrt{2}\}$.

Method	Medium				Hard				Multi-scale	
	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M	scale	dimen
(A) *Local features aggregation + re-ranking*										
HesAff-rSIFT-ASMK*+SP[35]	60.60	46.80	61.40	42.30	36.70	26.90	35.00	16.80	-	-
DELF-ASMK*+SP(GLDv1)[26, 29]	67.80	53.80	76.90	57.30	43.10	31.20	55.40	26.40	-	-
DELF-D2R-R-ASMK*+SP(GLDv1)[34]	76.00	64.00	80.20	59.70	52.40	38.10	58.60	29.40	-	-
R50-How-ASMK,n=2000[36]	79.40	65.80	81.60	61.80	56.90	38.90	62.40	33.70	-	-
FIRe(SfM-120k)[39]	**81.80**	**66.50**	**85.30**	**67.60**	**61.20**	**40.10**	**70.00**	**42.90**	7	-
(B) *Global features + Local feature re-ranking*										
R101-GeM+DSM [32]	65.30	47.60	77.40	52.80	39.20	23.20	56.20	25.00	-	-
R50-DELG(GLDv2-clean)[4]	78.30	67.20	85.70	69.60	57.90	43.60	71.00	45.70	3	2048
R101-DELG(GLDv2-clean)[4]	81.20	69.10	87.20	71.50	64.00	47.50	72.80	48.70	3	2048
R50-CVNet-Rerank(Top-400)[21]	**87.90**	80.70	90.50	82.40	75.60	65.10	80.20	67.30	3	2048
R101-CVNet-Rerank(Top-400)[21]	87.20	**81.90**	**91.20**	**83.80**	**75.90**	**67.40**	**81.10**	**69.30**	3	2048
(C) *Global features*										
R101-SOLAR(GLDv1)[25]	69.90	53.50	81.60	59.20	47.90	29.90	64.50	33.40	3	2048
R50-DOLG(GLDv2-clean)^r[42]	80.05	70.53	89.49	77.85	60.75	44.63	77.45	57.52	5	512
R101-DOLG(GLDv2-clean)^r[42]	**81.97**	72.43	90.11	80.24	**63.76**	48.28	78.20	61.33	5	512
R50-CVNet-Global(GLDv2-clean)[21]	81.00	72.60	88.80	79.00	62.10	50.20	76.50	60.20	3	2048
R101-CVNet-Global(GLDv2-clean)[21]	80.20	**74.00**	**90.30**	**80.60**	63.10	**53.70**	**79.10**	**62.20**	3	2048
R50-CFCD(GLDv2-clean)	**82.51**	**72.73**	**89.64**	**78.91**	**63.59**	**48.54**	**78.06**	**60.09**	3	2048
R101-CFCD(GLDv2-clean)	**84.08**	**74.66**	**91.03**	**82.18**	**67.80**	**54.10**	**81.21**	**65.51**	3	2048
R50-CFCD(GLDv2-clean)	**82.42**	**73.06**	**91.57**	**81.57**	**65.06**	**50.78**	**81.69**	**62.80**	5	512
R101-CFCD(GLDv2-clean)	**85.24**	**73.99**	**91.56**	**82.80**	**69.96**	**52.78**	**81.78**	**65.78**	5	512

Table 1. Results (% mAP) of different methods on $\mathcal{R}$Oxf(+1M) and $\mathcal{R}$Par(+1M) with Medium and Hard evaluation protocols. State-of-the-art performances are marked bold and our results are summarized in the bottom section. "$*$" means feature quantization is used. Methods superscripted with ^{r} are our re-implementations. "scale" and "dimen" are different scales and dimensions for global features. Our method belongs to the global features single pass group (C).

4.2. Results

In Tab.1, we divide the previous methods into three groups: (A) local features aggregation and re-ranking; (B) global features followed by local features re-ranking; and (C) global features. Our CFCD belongs to the group C. It can be observed that our solution consistently outperforms existing one-stage methods without additional computation overhead.

Comparison with One-stage State-of-the-art Methods. 1) Like methods in the global feature based group C, our method performs single-stage image retrieval with only the global feature. Due to the misreported results[1] in DOLG , we re-implement the R50/101-DOLG^r in the official configuration and achieve similar performance. R50/101-DOLG^r using local branch and orthogonal fusion module to combine both local and global information is still an excellent single-stage method. Our R50-CFCD with ResNet50 backbone even outperform R101-DOLG^r with ResNet101 backbone in all settings. Notably, our method R101-CFCD outperforms the R101-DOLG^r with a gain of up to 3.27% on $\mathcal{R}$Oxf-Medium, 1.45% on $\mathcal{R}$Par-Medium, 6.2% on $\mathcal{R}$Oxf-Hard and 3.58% on $\mathcal{R}$Par-Hard. Even with a large amount of distractors in the database, our R50-CFCD and R101-CFCD still outperform R50-DOLG^r and R101-DOLG^r by a large margin, respectively, which demonstrates the effectiveness of our method to exploit fine-grained local information. **2)** When compared with CVNet-Global, the proposed CFCD exhibits significantly superior performance in almost all dataset. These results exhibit excellent performance of our framework on single-stage image retrieval benchmarks. It should be noted that our method does not contain the expensive local attention module as in [42]. This suggests that infusing aligned local information into the final descriptor is a better option.

Comparison with Other Two-stage Methods. 1) In the local feature based group A, FIRe [39] is the current state-of-art local feature aggregation method and it outperforms R50-How-ASMK. Regardless of its complexity, our R50-CFCD outperforms it by 3.86% on Roxf-Hard and 11.69% on Rpar-Hard with the same ResNet50 backbone. **2)** In group B, another type of two-stage methods are based on

[1] https://github.com/feymanpriv/DOLG-paddle

#	MadaCos	Triplet	HNS	E	Medium		Hard	
					$\mathcal{R}$Oxf	$\mathcal{R}$Par	$\mathcal{R}$Oxf	$\mathcal{R}$Par
1				-	78.76	88.59	58.53	75.97
2	✓			-	81.86	90.80	64.38	80.21
3	✓	✓		50	82.04	91.23	64.84	81.68
4	✓	✓		0	78.18	89.58	58.29	75.79
5		✓	✓	50	79.96	89.54	60.46	77.61
6	✓	✓	✓	50	**82.42**	**91.57**	**65.06**	**81.69**

Table 2. Ablation study on different components. "MadaCos" means our median adaptive loss. "Triplet" means training with the triplet loss. "HNS" means the hard negative sampling strategy. "E" is the epoch when triplet loss is added for training.

Loss	ρ	Medium		Hard	
		$\mathcal{R}$Oxf	$\mathcal{R}$Par	$\mathcal{R}$Oxf	$\mathcal{R}$Par
ArcFace [8]	-	78.76	88.59	58.53	75.97
AdaCos[43]	-	75.88	87.44	56.62	73.99
MadaCos	0.01	80.44	89.46	62.36	77.57
	0.02	**81.78**	**90.60**	63.36	**79.94**
	0.03	80.87	89.62	62.30	78.46
	0.04	81.63	89.33	**64.90**	77.13
	0.05	81.72	89.51	63.18	77.51
CosFace [37]	0.02	71.01	82.39	50.28	68.7

Table 3. Ablation study on different ρ in MadaCos function.

the retrieve-and-rerank paradigm where the global retrieval is followed by a local feature re-ranking. Such methods exhibit superior performance owing to the nature of re-ranking. But our one-stage method (R50-CFCD) still outperforms the two-stage method (R101-DELG) on the both $\mathcal{R}$Oxf dataset and $\mathcal{R}$Par dataset by a significant margin. However, the re-ranking network of CVNet is trained with 1M images manually selected from 31k landmarks of the GLDv2-clean dataset. Since the authors of CVNet[21] do not upload their cleaned training data, it should be noted that comparisons in Tab.1 is not fair for our method. More comparisons with other two-stage methods are provided in the supplementary material.

Qualitative Analysis. We showcase top-5 retrieval results with a hard query image in Fig.5 of different methods. We can observe many false positives in the retrieval list of DOLG and CVNet-global, because its weak and implicit supervision on the local information is not robust enough when the query information is focused on a local patch. In contrast, more true positives can be recalled with only global features trained by our MadaCos loss. When additionally introducing triplet loss with the matching strategy to integrate local information into the global features, we obtain more robust retrieval results.

4.3. Ablation Studies

In this section, we conduct ablation experiments using the ResNet50 backbone to empirically validate the components of CFCD. We use 5 scales, $\{\frac{1}{2\sqrt{2}}, \frac{1}{2}, \frac{1}{\sqrt{2}}, 1, \sqrt{2}\}$ and set the dimension $\boldsymbol{d_g}$ of the global feature to 512.

Verification of Different Components. In Tab.2, we provide detailed ablation experimental results, verifying the contributions of three components in the framework by incrementally adding components to the baseline framework. For the baseline framework in the first row, we set the ArcFace margin m as 0.15 and scale s as 30 following DOLG, and train the model with ArcFace loss for 100 epoch. When the MadaCos loss is adopted to train for 100 epoch, mAP increases from 78.76% to 81.86% on $\mathcal{R}$Oxf-Medium and from 58.53% to 64.38% on $\mathcal{R}$Oxf-Hard. Then we introduce the coarse-to-fine framework to train the whole model with MadaCos and triplet losses, and observe that selecting local descriptors to discover patch-level matches between images helps to improve the overall performance, especially on hard cases. The mAP is improved from 64.38% to 64.84% on $\mathcal{R}$Oxf-Hard and from 80.21% to 81.68% on $\mathcal{R}$Par-Hard as in row 3. This indicates that aligning matching local descriptors according to visual patterns makes them more discriminative. The comparisons between 1st and 5th rows shows that the performance of ArcFace can be significantly improved by the coarse-to-fine framework. However, the comparisons between 3rd and 4th rows suggest that naively optimizing with total loss from the start leads to suboptimal performance, because during early training stage the local features are too premature for feature matching and may damage the global feature representation. In the last row, the performance is further improved when training with hard samples.

Loss Comparison for Global Feature Learning. The results of training with different anchor point ρ in softmax function are shown in Tab.3. Unlike the scale s and margin m of the ArcFace, the single anchor point ρ is the only manually tunable parameter in MadaCos. We simply adjust ρ from 0.01 to 0.05 to train the global descriptors for only 50 epochs, and the results are significantly improved, even surpassing R50-DOLG^r which is trained for 100 epochs.

As ρ increases, the mAP performance first increases and then decreases, with the best results at $\rho = 0.02$. Fig.6 illustrates the value change of scale s and margin m in MadaCos. As the training proceeds, the scale s and margin m gradually increase and then plateau out. In the last row of Tab.3, as we fix $s = 48.33, m = 0.33$ according to their converged values at the 50th epoch in Fig.6, MadaCos then degenerates into the initial CosFace, where the model performance drops sharply. This indicates that training with a large scale and margin at the beginning provides suboptimal supervision. Compared to existing parameter-free loss as AdaCos[43], our dynamically tuned scaling parameters are more efficient. More experiments on the face recogni-

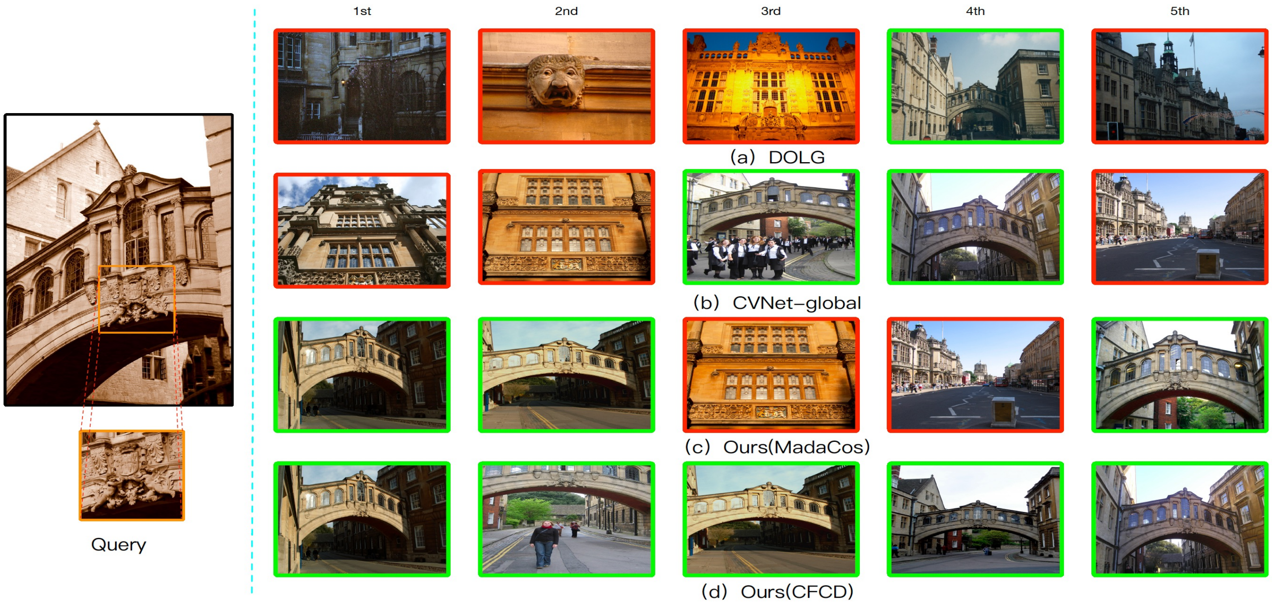


Figure 5. Demonstration of the top-5 retrieved results. The query on the left used as an input is generated by cropping only the part bounded by a orange box. On the right are the results of DOLG, CVNet-global and Ours(only Madacos loss and CFCD methods), which are shown from top to bottom. Green and red boxes denote positive and negative images, respectively.

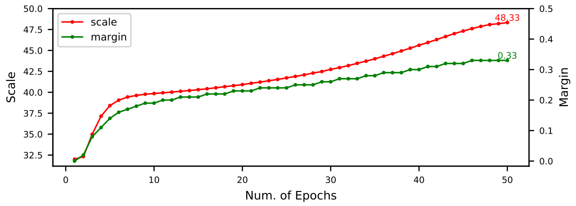


Figure 6. The values of the scale and margin of MadaCos during training with $\rho = 0.02$.

Config	Layer	Medium		Hard	
		$\mathcal{R}$Oxf	$\mathcal{R}$Par	$\mathcal{R}$Oxf	$\mathcal{R}$Par
w/o matching	$Res4$	80.97	91.22	62.37	80.51
matching	$Res4$	**82.42**	**91.57**	**65.06**	**81.69**

Table 4. Experiments of triplet loss with matching constraints. “w/o matching” means introducing triplet loss without matching constraints to the global descriptors before the whitening FC layer, “$Res4$” means selecting local descriptors based on Res4.

tion datasets are provided in the supplementary material.

Impact of Triplet Loss with Matching Constraints. We also provide experimental results to validate the impact of triplet loss with matching constraints. Our coarse-to-fine framework adopts MadaCos and triplet losses with different configurations to train model for 100 epochs, and the results are summarized in Tab.4 and Tab.5. In Tab.4, we can observe introducing the triplet loss with matching strategy to learn semantically salient region relations at $Res4$ improves the overall performance. We further explore the impact of controlling factor τ at $Res4$ in Tab.5 which indicates smaller τ increases the overall performance. This is because selecting matching background information may bring noise to the global features, and stricter matching constraints can help the global features integrate more discriminative local information.

$\tau(\%)$	Medium		Hard	
	$\mathcal{R}$Oxf	$\mathcal{R}$Par	$\mathcal{R}$Oxf	$\mathcal{R}$Par
30	**82.42**	**91.57**	65.06	**81.69**
50	82.02	91.41	**65.10**	81.46
70	81.88	91.35	64.94	81.38

Table 5. Experimental results of triplet loss with different controlling factors τ.

5. Conclusion

In this paper, we propose **C**oarse-to-**F**ine framework to learn **C**ompact and **D**iscriminative representation (CFCD), an end-to-end image retrieval framework which dynamically tunes the hyperparameters of its loss function progressively to strengthen supervision for improving intra-class compactness and leverages fine-grained semantic relations to infuse global feature with inter-class distinctiveness. The resulting framework is robust to local region variations as well as exhibits more potential to real-world applications due to its single-stage inference without additional computation overhead of local feature re-ranking. Extensive experiments demonstrates the effectiveness and efficiency of our method, which provides a practical solution to difficult retrieval tasks such as landmark recognition.

References

[1] Artem Babenko and Victor Lempitsky. Aggregating deep convolutional features for image retrieval. *arXiv preprint arXiv:1510.07493*, 2015. 2

[2] Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Bmvc*, volume 1, page 3, 2016. 2

[3] Michael M Bronstein, Alexander M Bronstein, Fabrice Michel, and Nikos Paragios. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3594–3601. IEEE, 2010. 2

[4] Bingyi Cao, Andre Araujo, and Jack Sim. Unifying deep local and global features for image search. In *European Conference on Computer Vision*, pages 726–743. Springer, 2020. 1, 2, 5, 6

[5] Wei Chen, Yu Liu, Weiping Wang, Erwin Bakker, Theodoros Georgiou, Paul Fieguth, Li Liu, and Michael S Lew. Deep image retrieval: A survey. *arXiv preprint arXiv:2101.11282*, 2021. 1

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[7] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *European Conference on Computer Vision*, pages 741–757. Springer, 2020. 2

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 1, 2, 3, 7

[9] Zelu Deng, Yujie Zhong, Sheng Guo, and Weilin Huang. Insclr: Improving instance retrieval with self-supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 516–524, 2022. 1

[10] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint detection and description of local features. *arXiv preprint arXiv:1905.03561*, 2019. 2

[11] Alaaeldin El-Nouby, Natalia Neverova, Ivan Laptev, and Hervé Jégou. Training vision transformers for image retrieval. *arXiv preprint arXiv:2102.05644*, 2021. 1

[12] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 1, 2

[13] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *Proc. European Conference on Computer Vision (ECCV)*, pages 241–257. Springer, 2016. 1

[14] Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Larlus. End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision*, 124(2):237–254, 2017. 2

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[16] Kun He, Yan Lu, and Stan Sclaroff. Local descriptors optimized for average precision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 596–605, 2018. 2

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[18] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3304–3311. IEEE, 2010. 2

[19] Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Pérez, and Cordelia Schmid. Aggregating local image descriptors into compact codes. *IEEE transactions on pattern analysis and machine intelligence*, 34(9):1704–1716, 2011. 2

[20] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3238–3247, 2020. 1

[21] Seongwon Lee, Hongje Seong, Suhyeon Lee, and Euntai Kim. Correlation verification for image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5374–5384, June 2022. 1, 5, 6, 7

[22] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017. 2, 3

[23] Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 360–368, 2017. 1

[24] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *European Conference on Computer Vision*, pages 681–699. Springer, 2020. 1

[25] Tony Ng, Vassileios Balntas, Yurun Tian, and Krystian Mikolajczyk. Solar: Second-order loss and attention for image retrieval. *Arxiv*, 2020. 6

[26] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*, pages 3456–3465, 2017. 2, 6

[27] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE conference on com-*

puter vision and pattern recognition, pages 1–8. IEEE, 2007. 5

[28] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008. 5

[29] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5706–5715, 2018. 5, 6

[30] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018. 2

[31] Jerome Revaud, Jon Almazán, Rafael S Rezende, and Cesar Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5107–5116, 2019. 2

[32] Oriane Siméoni, Yannis Avrithis, and Ondrej Chum. Local features and visual words emerge in activations. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11651–11660, 2019. 6

[33] Fuwen Tan, Jiangbo Yuan, and Vicente Ordonez. Instance-level image retrieval using reranking transformers. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[34] Marvin Teichmann, Andre Araujo, Menglong Zhu, and Jack Sim. Detect-to-retrieve: Efficient regional aggregation for image search. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5109–5118, 2019. 6

[35] Giorgos Tolias, Yannis Avrithis, and Hervé Jégou. Image search with selective match kernels: aggregation across single and multiple images. *International Journal of Computer Vision*, 116(3):247–261, 2016. 1, 2, 6

[36] Giorgos Tolias, Tomas Jenicek, and Ondřej Chum. Learning and aggregating deep local descriptors for instance-level recognition. In *European Conference on Computer Vision*, pages 460–477. Springer, 2020. 6

[37] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018. 2, 3, 7

[38] Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. Deep metric learning with angular loss. In *Proceedings of the IEEE international conference on computer vision*, pages 2593–2601, 2017. 2

[39] Weinzaepfel, Philippe and Lucas, Thomas and Larlus, Diane and Kalantidis, Yannis. Learning Super-Features for Image Retrieval. In *ICLR*, 2022. 2, 5, 6

[40] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584, 2020. 2, 3, 5

[41] Hui Wu, Min Wang, Wengang Zhou, Yang Hu, and Houqiang Li. Learning token-based representation for image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2703–2711, 2022. 1, 2

[42] Min Yang, Dongliang He, Miao Fan, Baorong Shi, Xuetong Xue, Fu Li, Errui Ding, and Jizhou Huang. Dolg: Single-stage image retrieval with deep orthogonal fusion of local and global features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11772–11781, 2021. 2, 3, 5, 6

[43] Xiao Zhang, Rui Zhao, Yu Qiao, Xiaogang Wang, and Hongsheng Li. Adacos: Adaptively scaling cosine logits for effectively learning deep face representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10823–10832, 2019. 3, 7

[44] Liang Zheng, Yi Yang, and Qi Tian. Sift meets cnn: A decade survey of instance retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(5):1224–1244, 2017. 2