Correlation Verification for Image Retrieval

Seongwon Lee Hongje Seong Suhyeon Lee Euntai Kim*
School of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea
{won4113, hjseong, hyeon93, etkim}@yonsei.ac.kr

Abstract

Geometric verification is considered a de facto solution for the re-ranking task in image retrieval. In this study, we propose a novel image retrieval re-ranking network named Correlation Verification Networks (CVNet). Our proposed network, comprising deeply stacked 4D convolutional layers, gradually compresses dense feature correlation into image similarity while learning diverse geometric matching patterns from various image pairs. To enable cross-scale matching, it builds feature pyramids and constructs cross-scale feature correlations within a single inference, replacing costly multi-scale inferences. In addition, we use curriculum learning with the hard negative mining and Hide-and-Seek strategy to handle hard samples without losing generality. Our proposed re-ranking network shows state-of-the-art performance on several retrieval benchmarks with a significant margin (+12.6% in mAP on $\mathcal{R}$Oxford-Hard+1M set) over state-of-the-art methods. The source code and models are available online: https://github.com/sungonce/CVNet.

1. Introduction

Image retrieval is a long-standing problem in computer vision. This task aims to sort a database of images based on their similarities to the given query image. For this task, global retrieval through global descriptor matching and geometric verification after local feature matching are mainly employed. These approaches typically comprise two primary components of the image retrieval framework that mutually complement one another. The global retrieval quickly performs a coarse retrieval across the database, and geometric verification re-ranks the coarse results by performing precise evaluation only on the potential candidates. Along with deep learning, image retrieval has also advanced significantly. In particular, several studies [8, 30, 41, 45, 46, 53] have been focused on extracting representative and distinctive features for global and local representations with deep learning. However, geometric verification after local feature matching still plays an essential role in the re-ranking task in image retrieval, despite its drawbacks. Owing to its *verify-after-matching* structure, geometric verification is performed based on only sparse and thresholded feature correspondence. Moreover, it is neither learnable nor differentiable and requires iterative optimization even during testing. In addition, geometric verification does not include a component that can handle multi-scale operation. Thus, several studies [8, 30, 32, 45] have attempted to solve the scale problem by repeating inference with the image pyramid to extract multi-scale local features. However, this is an extremely expensive process.

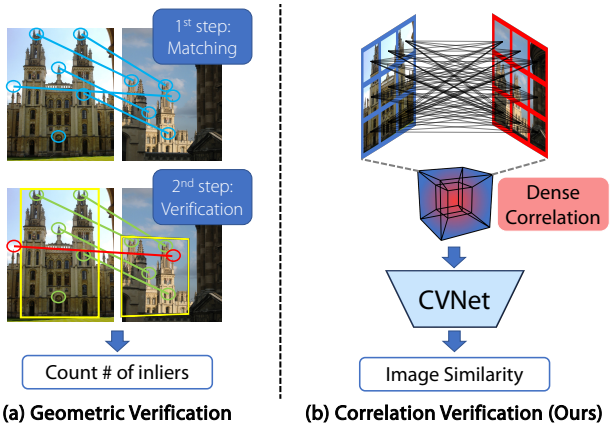

Figure 1. Novel image retrieval re-ranking method named correlation verification that *directly* predicts image similarity by leveraging dense feature correlation in a convolutional manner.

In this study, we propose an end-to-end learnable re-ranking network called *Correlation Verification Networks* (CVNet) to replace the role of geometric verification in a better way. The proposed network *directly* evaluates semantic and geometric relations by leveraging dense feature correlations in a convolutional manner. Following the successful architectural design of representative 2D convolutional neural networks (CNN), we design a 4D CNN with a pyramid structure of deeply stacked 4D convolution layers. It compresses the correlation between semantic cues into image similarity while learning diverse geometric matching patterns from a large number of image pairs. To ensure robustness even for large scale difference problems, it

*Corresponding author.

expands the single-scale feature to a feature pyramid for each image, forming cross-scale correlations between feature pyramids. This structure enables cross-scale matching with a single inference while replacing the multi-scale inference conventionally used in image retrieval. Our model does not require additional inference to extract local information; therefore the feature extraction latency, which significantly affects online retrieval time, is considerably reduced compared with other re-ranking methods. Similar to several computer vision problems, image retrieval suffers from the problem of hard samples. We address these challenges through curriculum learning using the hard negative mining and Hide-and-Seek [43] strategy in the training phase. This improves the overall performance by focusing on hard samples without losing generality in the case of normal ones. Our proposed re-ranking network shows state-of-the-art performance on several image retrieval benchmarks with a significant margin over several state-of-the-art methods. Our main contributions are as follows:

- We present Correlation Verification Networks (CVNet), which is a powerful re-ranking model that directly predicts the similarity of an image pair based on dense feature correlation.
- To replace expensive multi-scale inference, we construct cross-scale correlations within the model and perform cross-scale matching using a single inference.
- We propose curriculum learning using the hard negative mining and Hide-and-Seek strategy to handle hard samples without losing generality.
- The proposed model achieves new state-of-the-art performance on several image retrieval benchmarks: $\mathcal{R}$Oxford (+1M), $\mathcal{R}$Paris (+1M), and GLDv2-retrieval.

2. Related Work

Image retrieval. Over the past few decades, image retrieval [1, 8, 20, 21, 35, 36, 44, 46] has been one a primary focus of computer-vision studies. In pioneering research, handcrafted local features [6, 23] have been employed for global retrieval and re-ranking. A global retrieval with a global descriptor that aggregates handcrafted local features [19–21, 32, 33, 44] is performed first, and spatial verification [2, 32, 33] via local feature matching with RANSAC [12] is performed to re-rank putative retrieval results. Afterward, with the advancements in deep learning, global [1, 3, 4, 8, 13, 36, 48, 53] and local features [5, 8, 11, 24, 27, 28, 30, 54] extracted from deep-learning networks have replaced handcrafted features.

Although the techniques of global and local representations has progressed significantly, geometric verification remains a de facto solution for image retrieval re-ranking in both conventional [32, 33, 51] and recent studies [8, 30, 41, 46]. In a recent study, Reranking Transformers (RRT) [45] were proposed as a replacement for geometric verification by leveraging the transformer structure [49]. However, no significant improvement in performance was reported. In this study, we propose a novel re-ranking solution that exhibits powerful retrieval performance.

Diffusion / Query expansion. Among the re-ranking methods, several methods such as diffusion [9, 18] and query expansion [10, 36] exist that require additional expenses to traverse the entire database. However, because this study focuses on improving image matching for single pairs, we do not consider these re-ranking methods.

4D convolutional neural network. 4D convolution is a promising solution that has received considerable attention for tasks that require interpretation of the relationship between two images (*e.g.* visual dense correspondence prediction [22, 25, 38, 52] and few-shot segmentation [26]). The primary difference between the aforementioned tasks and image retrieval is that the former aims for a 2D (single image side) [26] or 4D (both image sides) [25, 25, 52] dense output, whereas the latter requires a single similarity value. Therefore, in this study, we propose a novel structure that gradually compresses the 4D feature correlation through deeply stacked 4D convolution layers.

Hide-and-Seek. Hide-and-Seek [43] is an augmentation technique that has been proposed to improve object localization performance in weakly supervised fields. To address the drawback that the network focuses only on the most salient areas, a few random patches of the image are masked to induce the network to make robust predictions despite having visual access only to less salient areas. We found that the Hide-and-Seek approach could improve the image retrieval performance by enabling accurate matching even on hard samples, such as those involving occlusion or truncation. In this study, we apply Hide-and-Seek to our model in a curriculum manner to ensure robustness when handling hard samples without losing generality.

3. Global Backbone Network (CVNet-Global)

In this section, we introduce our proposed global backbone network named CVNet-Global. An overview of CVNet-Global is shown in Fig. 2. Our proposed global backbone network, that takes a single image $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$ as the input, is used to extract the global descriptor $\mathbf{d}_g \in \mathbb{R}^{C_g}$ for global image retrieval and local feature map $\mathbf{F} \in \mathbb{R}^{C_l\times H_l\times W_l}$ for the re-ranking phase. We adopt multi-objective loss [7] that jointly optimizes the classification loss and contrastive loss to induce the network to learn more distinctive and robust global and local representations.

3.1. Structure

Inspired by the momentum-contrastive structure of MoCo [15], we build two networks: the global backbone network f and its momentum network $\bar{f}$. These two networks are based on ResNet [16]. f_i denotes ith *ResBlock*.

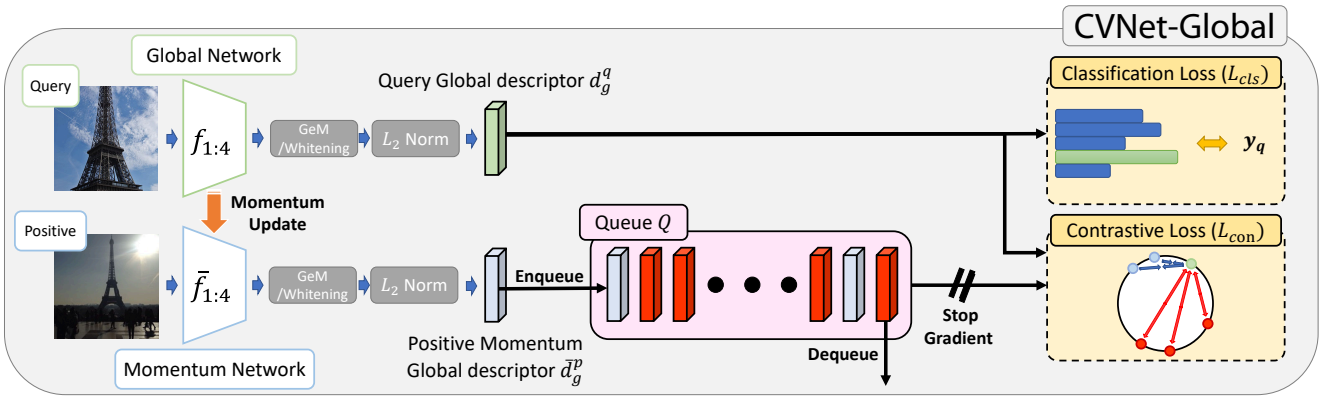

Figure 2. Illustration of the proposed Global backbone network (CVNet-Global) and its training objective. The network has two objectives: classification loss and contrastive loss. To utilize several samples without a computational burden in contrastive learning, momentum network and queue structure are adopted from MoCo [15]. The combination of these objectives enables the network to learn intra-class variability and inter-class distinctiveness, which is required for image retrieval task.

Global Average Pooling is replaced with learnable GeM pooling [35] with power initialized to 3.0, and a whitening FC layer [14] and L2-normalization are added after the pooling layer. We build a queue $\mathbf{Q} \in \{\bar{\mathbf{d}}_g^i\}_{i=1}^K$, to save momentum global descriptors for each iteration and utilize them as contrastive samples.

3.2. Training Objective

Classification loss. At each iteration, the query image $\mathbf{I}_q$ is fed into the global network f to compute the query global descriptor $\mathbf{d}_g^q$. With $\mathbf{d}_g^q$, CurricularFace [17]-margined classification loss $\mathcal{L}_{cls}$ is computed as

$$\mathcal{L}_{cls} = -\log \frac{\exp(\mathcal{C}(\mathbf{W}_{y_g}{}^T \mathbf{d}_g^q, 1)/\tau)}{\sum_{i=1}^{N} \exp(\mathcal{C}(\mathbf{W}_{y_i}{}^T \mathbf{d}_g^q, \mathbb{1}_q^i)/\tau)}, \quad (1)$$

where $\mathbf{W}$ is the class weight, τ is the scale parameter, y_g is the ground-truth class, and $\mathbb{1}_q^i$ is an indicator that shows whether the ith class y_i is identical to y_g. $\mathcal{C}$ is a function that adds a CurricularFace margin to query-positive cosine similarity.

Momentum contrastive loss. At each iteration, a positive image $\mathbf{I}_p$ with the same label as the query image $\mathbf{I}_q$ is sampled and fed into the momentum network $\bar{f}$ to compute the positive momentum global descriptor $\bar{\mathbf{d}}_g^p$. The descriptor $\bar{\mathbf{d}}_g^p$ is updated to queue $\mathbf{Q}$ while dequeuing the last element of the queue. Then, queue $\mathbf{Q}$ holds at least one momentum sample with the same label as the query including $\bar{\mathbf{d}}_g^p$. Thus we use the CurriculurFace-margined momentum contrastive loss $\mathcal{L}_{con}$:

$$\mathcal{L}_{con} = \frac{-1}{|P(q)|} \sum_{p \in P(q)} \log \frac{\exp\left(\bar{\mathcal{C}}\left(\mathbf{d}_g^q \cdot \bar{\mathbf{d}}_g^p, 1\right)/\tau\right)}{\sum_{i \in \{p\} \bigcup N(q)} \exp\left(\bar{\mathcal{C}}\left(\mathbf{d}_g^q \cdot \bar{\mathbf{d}}_g^i, \mathbb{1}_q^i\right)/\tau\right)}, \quad (2)$$

where $\bar{\mathcal{C}}$ is identical to $\mathcal{C}$, but updates its moving average parameter separately with $\mathcal{C}$. $P(q)$ and $N(q)$ are the in-queue positive and negative set, respectively.

Total loss. Finally, the total loss of our global backbone network $\mathcal{L}_g$ is the weighted sum of the classification loss $\mathcal{L}_{cls}$ and contrastive loss $\mathcal{L}_{con}$:

$$\mathcal{L}_g = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{con}\mathcal{L}_{con}. \quad (3)$$

Note that, optimizer only updates the global backbone network f. The momentum network $\bar{f}$ is momentum updated with a momentum of η.

4. Re-Ranking Network (CVNet-Rerank)

In this section, we introduce our proposed re-ranking network, named CVNet-Rerank. An overview of CVNet-Rerank is shown in Fig. 3. Our proposed re-ranking network, which takes a pair of local feature maps $(\mathbf{F}_q, \mathbf{F}_k)$ of images $(\mathbf{I}_q, \mathbf{I}_k)$ as input, is used to predict the similarity $s_l^{q,k} \in \mathbb{R}^1$ between two images. It subsequently re-ranks the global image retrieval results based on the results of the predicted similarity. The local feature maps $(\mathbf{F}_q, \mathbf{F}_k)$ are extracted from the intermediate layer of the global backbone network f, that is fully trained and frozen. Representative 2D CNN architectures (*e.g.* VGG [42] and ResNet [16]) stack several 2D convolutional layers, followed by spatial-dimensional down-sampling to capture diverse level features in an image and compress it to fine-grained information. Inspired by the aforementioned structure, the proposed re-ranking network gradually compresses the feature correlation with deeply stacked 4D convolution layers and predicts the image similarity using the classifier.

4.1. Cross-scale Correlation Construction

Because image retrieval must be robust for scale difference, several image retrieval methods that use local features built a multi-scale local feature set through multiple inferences using an image pyramid. Here, following [25], we expand the extracted feature map to a multi-scale feature pyramid to capture semantic cues from different scales inside the model, thus avoiding the expensive task

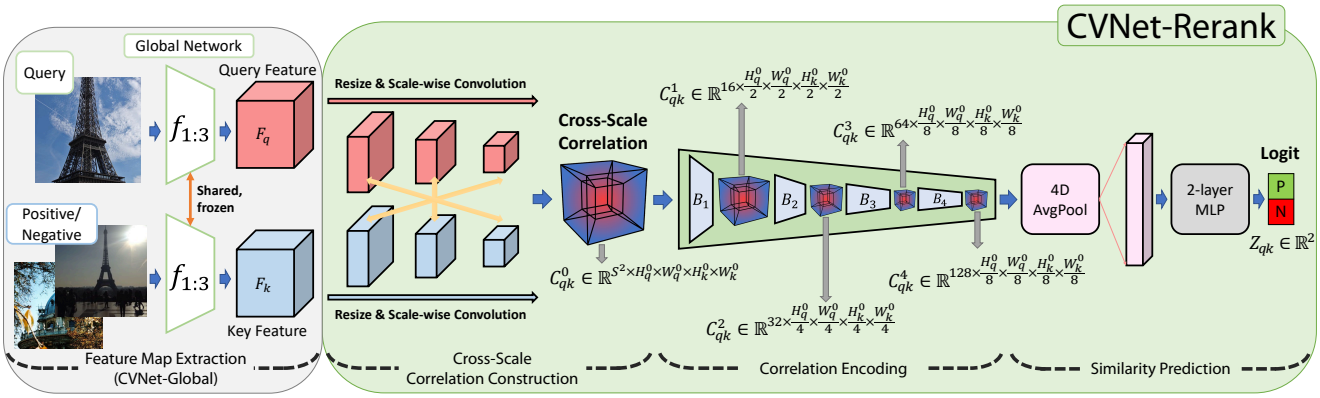

Figure 3. Illustration of the proposed Re-ranking network (CVNet-Rerank). The proposed network takes pair of feature maps extracted from the trained CVNet-Global model as input, constructs a cross-scale feature correlation, and gradually compresses it to image similarity of a pair with deeply stacked 4D convolution layers.

of multi-scale inference. Given a pair of query and key images $\mathbf{I}_q, \mathbf{I}_k \in \mathbb{R}^{3\times H\times W}$, we extract the local feature maps $\mathbf{F}_q, \mathbf{F}_k \in \mathbb{R}^{C_l\times H_l\times W_l}$ using the global backbone network f. After feature extraction, we construct a feature pyramid $\{\mathbf{F}^s\}_{s=1}^{S}$,where S is the number of scales, by repeatedly resizing the extracted feature map F with a scaling factor of $1/\sqrt{2}$. Each level of the feature pyramid passes the scale-wise 3×3 convolution layer, thereby reducing the channel dimension of each layer to C_l' to capture semantic information with diverse receptive field sizes while reducing the memory footprint of our image retrieval framework. With the constructed query feature pyramid $\{\mathbf{F}_q^s\}_{s=1}^{S}$ and key feature pyramid $\{\mathbf{F}_k^s\}_{s=1}^{S}$, we compute a 4-dimensional cross-scale correlation set $\{\mathbf{C}_{qk}^{s_q,s_k}\}_{(s_q,s_k)=(1,1)}^{(S,S)}$ of size S^2 using cosine similarity and ReLU function:

$$\mathbf{C}_{qk}^{s_q,s_k}(\mathbf{p}_q,\mathbf{p}_k)=\text{ReLU}\left(\frac{\mathbf{F}_q^{s_q}(\mathbf{p}_q)\cdot\mathbf{F}_k^{s_k}(\mathbf{p}_k)}{\left\|\mathbf{F}_q^{s_q}(\mathbf{p}_q)\right\|\left\|\mathbf{F}_k^{s_k}(\mathbf{p}_k)\right\|}\right), \quad (4)$$

where $\mathbf{p}_q$ and $\mathbf{p}_k$ are the pixel positions in each feature map. Finally, we interpolate all the correlations to obtain the original feature resolution $H_l \times W_l$ for each image side, stack all the correlations, and construct a cross-scale correlation set $\mathbf{C}_{qk}^0 \in \mathbb{R}^{S^2\times H_l\times W_l\times H_l\times W_l}$.

4.2. 4D Correlation Encoder

Our correlation encoder takes the cross-scale correlation set $\mathbf{C}_{qk}^0 \in \mathbb{R}^{S^2\times H_l\times W_l\times H_l\times W_l}$ and gradually compresses it into a binary class logit $\mathbf{Z}_{qk}=\{z_0,z_1\}\in\mathbb{R}^2$. We construct our encoder with a sequence of 4D convolution blocks, followed by a global average pooling layer and a 2-layer MLP classifier. Except for the last 4D convolution block, the remaining blocks perform spatial dimension down-sampling by constructing each last convolutional layer as a stride convolution. Naïve 4D convolution is computationally intensive and, therefore, unsuitable for online re-ranking. Using the knowledge taken from findings of previous studies, we

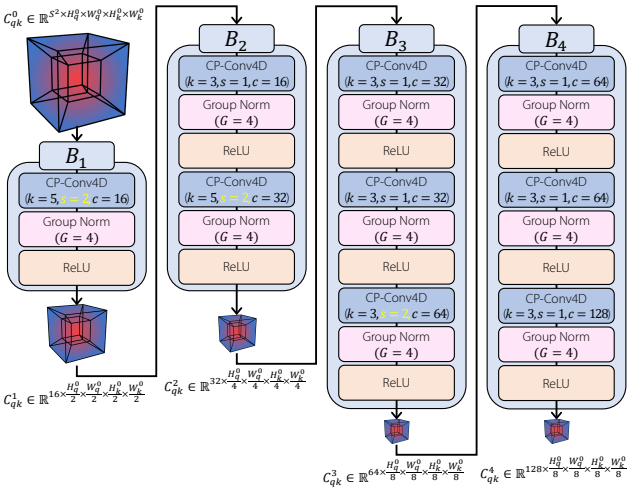

Figure 4. The detailed structure of the proposed 4D correlation Encoder. The proposed encoder structure gradually compresses the cross-scale correlation into a fine-grained correlation cue.

adopt a center-pivot 4D convolution [26] to reduce the burden of using high-dimensional kernels and enable real-time image re-ranking. With this pyramid structure of 4D convolution, the cross-scale feature correlation set is encoded as a fine-grained correlation cue $\mathbf{C}_{qk}^{1:4}$. It is subsequently converted into a class logit $\mathbf{Z}_{qk}$ through spatial dimension average pooling and a binary classifier.

4.3. Training Objective

Our re-ranking network is trained to minimize the cross-entropy loss for query and key pair (q,k):

$$\mathcal{L}_r^{qk}=\mathbf{CE}(\mathbf{Softmax}(\mathbf{Z}_{qk}),\mathbb{1}_q^k). \quad (5)$$

We symmetrically convert the loss $\mathcal{L}_r^{qk}$ to $\mathcal{L}_r^{kq}$ by reversing the query-key position. Afterward, we apply them to positive p and negative key samples n, respectively. The final loss for our re-ranking network is constructed as follows:

$$\mathcal{L}_r=\left(\mathcal{L}_r^{qp}+\mathcal{L}_r^{pq}+\mathcal{L}_r^{qn}+\mathcal{L}_r^{nq}\right)/4. \quad (6)$$

Figure 5. Examples of the query and hard negative samples of the GLDv2-clean dataset. These pairs look similar at first glance, but a closer look reveals several differences.

4.4. Training with Hard Samples

Because image re-ranking is performed on images that look similar at first glance, it must be robust against hard samples. Thus, we propose a method to train a network by focusing on hard samples through hard negative mining and Hide-and-Seek augmentation. Although hard samples are beneficial for model training, a possibility of losing generality in the case of normal samples exists. Carefully considering this concern, we apply hard negative mining and Hide-and-Seek augmentation in a curriculum learning manner to train the re-ranking network to make more accurate predictions without losing generality in the case of normal ones while concentrating on hard samples.

Hard negative mining. We selected hard-negative samples with help of trained global descriptors. For every sample in the training dataset, the top 10 negatives are selected in order of the highest global descriptor matching score. Example results of hard negative mining are shown in Fig. 5.

Hide-and-Seek. Similar to several computer vision studies, occlusion is a primary obstacle in image retrieval tasks. To solve this problem, we apply Hide-and-Seek [43] augmentation to synthetically generate matching situations that involve occlusions. In the original Hide-and-Seek method, the input image is divided into grids, and probabilistic deactivation is applied to each grid section. Similarly, we randomly deactivate each pixel value from each input feature map. This can have an effect similar to that of applying occlusion to the receptive field of the original image that corresponds to one pixel in the feature map. This concept is illustrated in Fig. 6.

Curriculum learning. To prevent hard samples from interfering with early learning, we apply hard negative mining and Hide-and-Seek in a curriculum learning manner. Instead of focusing on hard negatives from the outset, the rate of selecting hard negatives r_H and the probability of Hide-and-Seek augmentation p_{has} gradually increase as learning progresses. This curriculum learning helps the network to retain its generality to ensure that it consistently performs well even when the re-ranking range is extended.

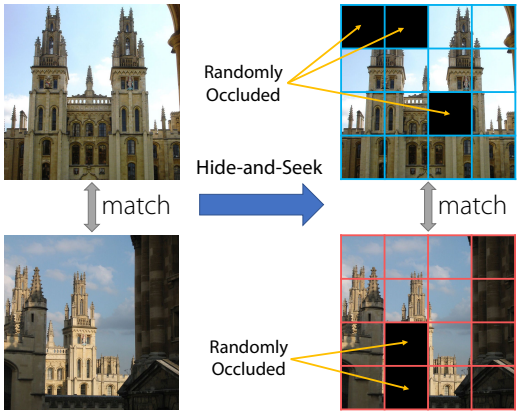

Figure 6. With Hide-and-Seek, the re-ranking network can effectively learn hard-matching cases by randomly hiding parts of matching pairs to give images an occlusion-like effect.

5. Experiments

5.1. Implementation Details

Common setting. Our proposed CVNet is implemented using PyTorch [31]. We use the 'clean' subset [55] of Google Landmarks dataset v2 (1.58M images from 81k landmarks) [50] as a training set. The input image is augmented with random cropping/aspect ratio distortion and resized to 512×512. We use an SGD optimizer with a momentum of 0.9 and use cosine learning rate scheduling.

Global backbone network. We use ResNet-50 (R50) and ResNet-101 (R101) as the encoder of global backbone networks with ImageNet [39] pre-trained weights, whereas ResNet-50 is used for ablation studies. We use a Shuffling Batch Normalization [15], global descriptor size of 2048, and a queue size of 73,728. We set the τ to $1/30$, m to 0.15, η to 0.999, and λ_{cls} and λ_{con} to 0.5. The global model is trained for 25 epochs (39.5M steps) for the training dataset, using a learning rate of 0.005625, and a batch size of 144.

Re-ranking network. For cross-scale correlation construction, we use $S = 3$ scales (*i.e.* $\{1/2, 1/\sqrt{2}, 1\}$). We extract the feature map $\mathbf{F}$ from the f_3 output and compress its channel dimension to $C_l' = 256$. Our training set contains various views of landmarks, including cases with no overlap. To avoid query-positive non-overlapping, we select verified match pairs for each class with help of deep local features [30] and exclude only those classes with a number of verified match pairs. Please see the supplementary material for a more detailed explanation of the data selection and sampling process used for the CVNet-Rerank. Finally, we select 1M images from 31k landmarks, and the proposed re-ranking model is trained for 200 epochs (6.3M steps) for all classes, using a learning rate of 0.00375 and a batch size of 96. r_H and p_{has} linearly increase from 0.2 to 1.0 and from 0 to 0.2 while training, respectively.

Feature extraction and matching. For global descriptor extraction, we follow the convention of previous stud-

Method	Medium				Hard				Multi-scale	
	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M	global	local
(A) Local feature aggregation (+ Local feature re-ranking)										
DELF-ASMK*+SP [30, 34]	67.8	53.8	76.9	57.3	43.1	31.2	55.4	26.4	-	7
DELF-D2R-R-ASMK* (GLDv1) [46]	73.3	61.0	80.7	60.2	47.6	33.6	61.3	29.9	-	7
+ SP (Rerank Top-100) [46]	76.0	64.0	80.2	59.7	52.4	38.1	58.6	29.4	-	7
R50-How-ASMK,n=2000 [47]	79.4	65.8	81.6	61.8	56.9	38.9	62.4	33.7	-	7
(B) Global features (+ Local feature re-ranking)										
R101-GeM$^{\uparrow}$ [36, 41]	65.3	46.1	77.3	52.6	39.6	22.2	56.6	24.8	3	-
+DSM (Rerank Top-100) [41]	65.3	47.6	77.4	52.8	39.2	23.2	56.2	25.0	3	3
R101-GeM-AP (GLDv1) [37]	66.3	-	80.2	-	42.5	-	60.8	-	1	-
R101-GeM+SOLAR (GLDv1) [29]	69.9	53.5	81.6	59.2	47.9	29.9	65.5	33.4	3	-
R50-DELG (Global-only, GLDv2-clean) [8]	73.6	60.6	85.7	68.6	51.0	32.7	71.5	44.4	3	-
+ GV (Rerank Top-100) [8]	78.3	67.2	85.7	69.6	57.9	43.6	71.0	45.7	3	7
+ GV (Rerank Top-200) [8, 45]	79.2	68.2	85.5	69.6	57.5	42.9	67.2	44.5	3	7
+ RRT (Rerank Top-100) [45]	78.1	67.0	86.7	69.8	60.2	44.1	75.1	49.4	3	7
+ RRT (Rerank Top-200) [45]	79.5	68.6	87.8	71.5	62.5	46.3	77.1	52.3	3	7
R101-DELG (Global-only, GLDv2-clean) [8]	76.3	63.7	86.6	70.6	55.6	37.5	72.4	46.9	3	-
+ GV (Rerank Top-100) [8]	81.2	69.1	87.2	71.5	64.0	47.5	72.8	48.7	3	7
+ RRT (Rerank Top-100) [8]	79.9	-	87.6	-	64.1	-	76.1	-	3	7
+ SuperGlue (Rerank Top-100) [8, 40]	79.7	-	87.1	-	62.1	-	71.5	-	3	7
R50-DOLG (GLDv2-clean) [53]	80.5	76.6	89.8	80.8	58.8	52.2	77.7	62.8	5	
R101-DOLG (GLDv2-clean) [53]	81.5	77.4	91.0	83.3	61.1	54.8	80.3	66.7	5	
(C) Ours										
R50-CVNet-Global (GLDv2-clean)	81.0	72.6	88.8	79.0	62.1	50.2	76.5	60.2	3	-
+ CVNet-Rerank (Rerank Top-100)	86.1	77.6	89.4	79.9	72.8	61.1	78.6	63.9	3	1
+ CVNet-Rerank (Rerank Top-200)	87.2	78.9	90.0	81.2	74.5	62.9	79.5	66.0	3	1
+ CVNet-Rerank (Rerank Top-400)	**87.9**	**80.7**	**90.5**	**82.4**	**75.6**	**65.1**	**80.2**	**67.3**	3	1
R101-CVNet-Global (GLDv2-clean)	80.2	74.0	90.3	80.6	63.1	53.7	79.1	62.2	3	-
+ CVNet-Rerank (Rerank Top-100)	85.6	79.6	90.6	81.5	72.9	64.5	80.4	66.2	3	1
+ CVNet-Rerank (Rerank Top-200)	86.4	81.0	91.1	82.7	74.6	66.6	81.0	68.0	3	1
+ CVNet-Rerank (Rerank Top-400)	**87.2**	**81.9**	**91.2**	**83.8**	**75.9**	**67.4**	**81.1**	**69.3**	3	1

Table 1. **Comparison with state-of-the-art methods.** Performance comparison on $\mathcal{R}$Oxf/$\mathcal{R}$Par and 1M-added experiments (referred to as +1M) with Medium and Hard evaluation protocols. The proposed image retrieval framework outperforms state-of-the-art image retrieval methods by a large margin for every measure. The best and second-best scores are presented as **boldfaced** and underlined text, respectively.

ies [8, 13, 30, 36, 45]. We extract global descriptors of three scales: $\{1/\sqrt{2}, 1, \sqrt{2}\}$. The final global descriptor is calculated by L2-normalizing the average of the three descriptors. During the re-ranking process, the final ranking is decided based on the final score $s_g + \alpha s_r$, where s_g is the cosine similarity of the global descriptors, s_r is the output score of the re-ranking network and α is the weight for s_r. As in previous studies [8, 29, 37, 46], the weight α is tuned in $\mathcal{R}$Oxf/$\mathcal{R}$Par and fixed for its large-scale experiment and GLDv2-retrieval test. Finally, we set the α to 0.5.

5.2. Evaluation Benchmarks

We primarily evaluate our model on $\mathcal{R}$Oxford5k [32, 34] (referred to as $\mathcal{R}$Oxf) and $\mathcal{R}$Paris6k [33, 34] (referred to as $\mathcal{R}$Par) datasets. Both datasets comprise 70 queries and 4933 and 6322 database images, respectively. In addition, an $\mathcal{R}$1M distractor set [34] is used for measuring the large-scale retrieval performance. Performance is measured using a mean Average Precision (mAP) metric. Additionally, we evaluate our model on the instance-level large-scale image retrieval task of the Google Landmarks dataset v2 [50] (referred to as GLDv2-retrieval). The GLDv2-retrieval comprises 750 test query images and 762k database images. In this task, performance is evaluated using a mean Average Precision@100 (mAP@100) metric.

5.3. Results

In this section, we compare our model with state-of-the-art image retrieval methods.

Comparison with state-of-the-art methods. (Tab. 1, Tab. 2) Tab. 1 shows a comparison between results of the proposed model and state-of-the-art image retrieval methods on $\mathcal{R}$Oxf and $\mathcal{R}$Par, and their +1M experiments. For all settings, the proposed CVNet outperforms the state-of-the-art methods. Our global model shows performance comparable to the state-of-the-art methods without additional modules, and our proposed re-ranking network exhibits superior performance without using expensive multi-scale inference. Because of the nature of re-ranking, the proposed model exhibits significantly superior performance in the difficult dataset ($\mathcal{R}$Oxf), for the difficult protocol (Hard), when a large number of images interfere (+1M). Our re-ranking method yields an improvement of up to 14.9% (R50-$\mathcal{R}$Oxf-Hard+1M), which is significantly higher than any of the state-of-the-art methods. In addition, the proposed method performs well without loss of generality even when the number of re-ranking samples increases. Tab. 2 compares CVNet with the results of the previous study's GLDv2-retrieval test. Even in this comparison, our proposed CVNet outperforms all state-of-the-art methods.

Method	mAP@100
DELF-R-ASMK*+SP [46]	18.8
R101-GeM+ArcFace [50]	20.7
R101-GeM+CosFace [55]	21.4
R50-DELG (GLDv2-clean) [8]	24.1
+ GV (Rerank Top-100) [8]	24.3
R101-DELG (GLDv2-clean) [8]	26.0
+ GV (Rerank Top-100) [8]	26.8
R50-CVNet-Global (Ours)	30.2
+ CVNet-Rerank (Rerank Top-100) (Ours)	**32.4**
R101-CVNet-Global (Ours)	32.5
+ CVNet-Rerank (Rerank Top-100) (Ours)	**34.9**

Table 2. **GLDv2-retrieval evaluation.** The result on the test split of the GLDv2-retrieval. The best scores are presented as **boldfaced** text for each ResNet backbone.

Comparison with other re-ranking methods. (Tab. 3) For a fair comparison, we attach the local branch of the DELG [8] to our global backbone to learn the local DELG features. With these learned local features, we reproduce two re-ranking methods: geometric verification (GV) and Reranking Transformer [45]. Details of the reproduction are provided in the supplementary material. While GV exhibits moderate performance improvement, RRT exhibits a decrease in performance in some sets, despite using the official code and setting. Our proposed method surpasses both methods by a large margin for all the measures.

5.4. Ablation Experiments

In this section, we present the core ablation results in Tab. 4. Please refer to the supplementary material for a detailed explanation of this and additional ablation studies.

Cross-scale correlation (Tab. 4a). We conduct an ablation study using cross-scale correlation construction to demonstrate its efficacy. The cross-scale correlation boosts the re-ranking performance, especially in hard protocols that include large-scale differences.

Hard negative mining and Hide-and-Seek (Tab. 4b). Our results demonstrate the effects of hard negative mining and Hide-and-Seek augmentation. When learning is performed only with random negatives, the network lost its distinguishing power and fails to re-rank. Considering the nature of re-ranking, that the process of re-ranking primarily encounters hard samples during testing, learning that focuses on hard negatives considerably improves performance. Hide-and-Seek augmentation also improves the overall performance by inducing the network to be robust against hard situations.

Loss comparison for the CVNet-Global (Tab. 4c). For the global backbone network, instead of using either the classification or contrastive loss, it is found that using both simultaneously results in overall improved performance.

Quantization (Tab. 4d). To reduce the memory footprint, we conduct an experiment by quantizing the multi-scale features stored in 32 bits to 8 bits. While this quantization reduces the memory footprint by 1/4, it hardly diminishes the overall performance.

#	Method	Medium				Hard			
		$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M
0	CVNet-Global	81.0	72.6	88.8	79.0	62.1	50.2	76.5	60.2
100	GV† [8]	82.2	74.0	89.0	79.3	64.2	51.9	77.1	60.8
	RRT† [45]	82.2	72.4	88.8	78.8	66.1	52.3	75.6	57.4
	CVNet-Rerank	**86.1**	**77.6**	**89.4**	**79.9**	**72.8**	**61.1**	**78.6**	**63.9**
200	GV† [8]	82.7	74.8	89.1	79.4	65.0	52.3	77.5	60.8
	RRT† [45]	82.1	71.6	88.7	77.9	66.0	51.3	75.2	53.5
	CVNet-Rerank	**87.2**	**78.9**	**90.0**	**81.2**	**74.5**	**62.9**	**79.5**	**66.0**
400	GV† [8]	82.5	74.8	89.1	79.5	63.8	52.1	77.5	61.1
	RRT† [45]	81.7	71.2	88.2	75.2	65.2	50.4	74.8	49.9
	CVNet-Rerank	**87.9**	**80.7**	**90.5**	**82.4**	**75.6**	**65.1**	**80.2**	**67.3**

Table 3. **Comparison with other re-ranking methods.** Geometric Verification (GV) and Reranking Transformers (RRT) are reproduced based on our R50-CVNet-Global. † indicates reproduced. # is the number of samples that is re-ranked and the best and second-best scores are presented as **boldfaced** and underlined text, respectively.

Extraction latency and memory footprint (Tab. 4e). Our feature extraction in the re-ranking process requires only a single inference, which is included in the process of extracting the global descriptor. Therefore, it has the lowest extraction latency time among the reproduced re-ranking methods. The memory footprint of the original model is large because of its dense nature. Thus, we attempt to reduce it with quantization (CVNet^{Q}). Through channel reduction and quantization, we achieve a memory footprint similar to that of re-ranking methods using sparse features while significantly improving the performance. Latency and matching time are measured on NVIDIA TITAN RTX GPU and i5-9600K CPU, for squared images of side 512. The time measured in the CPU is marked with an $*$.

6. Discussion

Qualitative results. Examples of our re-ranking results are provided in Fig. 7. Despite technological advances, global descriptor matching is easily fooled by similar-looking negative images and has difficulty finding occluded or truncated positives, even more so at different scales. Our re-ranking network can respond to scale changes owing to cross-scale correlation and has been trained to be robust in situations involving challenges such as occlusion. Consequently, our re-ranking network shows robust final retrieval results by boosting the ranks of positives even in cases where global descriptors are misjudged. Additional qualitative results are provided in the supplementary material.

Limitations and future work. Although our proposed re-ranking method has significant potential, it has shortcomings in terms of speed and memory, owing to its dense nature. To solve this problem, we apply kernel sparsification, channel reduction, and quantization to bring them up to an appropriate level, but the proposed method still re-

#	CSC	Medium				Hard			
		$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M
0		81.0	72.6	88.8	79.0	62.1	50.2	76.5	60.2
100		84.9	76.1	88.8	79.3	69.9	57.4	76.3	61.1
	✓	**86.1**	**77.6**	**89.4**	**79.9**	**72.8**	**61.1**	**78.6**	**63.9**
200		85.3	76.7	88.9	79.5	70.5	58.3	76.3	61.5
	✓	**87.2**	**78.9**	**90.0**	**81.2**	**74.5**	**62.9**	**79.5**	**66.0**
400		85.5	77.6	89.0	79.7	70.7	59.3	76.4	61.6
	✓	**87.9**	**80.7**	**90.5**	**82.4**	**75.6**	**65.1**	**80.2**	**67.3**

(a) **Cross-Scale Correlation.**

#	HNM	HaS	Medium				Hard			
			$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M
0			81.0	72.6	88.8	79.0	62.1	50.2	76.5	60.2
100			81.4	72.7	88.8	79.0	62.4	50.3	76.4	60.2
	✓		85.8	77.5	89.3	79.9	71.6	60.5	78.1	63.7
	✓	✓	**86.1**	**77.6**	**89.4**	**79.9**	**72.8**	**61.1**	**78.6**	**63.9**
200			81.3	72.6	88.7	78.9	62.5	50.2	76.5	60.2
	✓		86.9	78.7	89.7	81.0	73.4	62.1	78.6	65.6
	✓	✓	**87.2**	**78.9**	**90.0**	**81.2**	**74.5**	**62.9**	**79.5**	**66.0**
400			81.2	72.5	88.8	78.9	62.5	50.2	76.9	60.4
	✓		87.5	80.3	89.9	82.0	74.2	64.3	78.9	66.4
	✓	✓	**87.9**	**80.7**	**90.5**	**82.4**	**75.6**	**65.1**	**80.2**	**67.3**

(b) **Hard Negative Mining (HNM) and Hide-and-Seek (HaS).**

$\mathcal{L}_{cls}$	$\mathcal{L}_{con}$	Medium				Hard			
		$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M
✓		78.0	69.4	**89.8**	77.3	57.1	42.9	**78.4**	56.9
	✓	80.1	**73.5**	87.7	76.2	**62.2**	**51.9**	74.0	56.4
✓	✓	**81.0**	72.6	88.8	**79.0**	62.1	50.2	76.5	**60.2**

(c) **Loss Comparison of CVNet-Global.**

#	8-bit quant	Medium				Hard			
		$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M
0		81.0	72.6	88.8	79.0	62.1	50.2	76.5	60.2
100		**86.1**	**77.6**	**89.4**	**79.9**	**72.8**	**61.1**	**78.6**	**63.9**
	✓	**86.1**	**77.6**	**89.4**	**79.9**	**72.8**	**61.1**	**78.6**	**63.9**
200		**87.2**	**78.9**	**90.0**	**81.2**	**74.5**	**62.9**	**79.5**	**66.0**
	✓	**87.2**	**78.9**	**90.0**	**81.2**	**74.5**	62.8	**79.5**	**66.0**
400		**87.9**	**80.7**	**90.5**	**82.4**	**75.6**	**65.1**	**80.2**	**67.3**
	✓	**87.9**	80.6	**90.5**	**82.4**	75.5	**65.1**	**80.2**	**67.3**

(d) **8-bit Quantization.**

Method	Multi-scale		Extraction latency (ms)			Matching time (ms)	Memory (GB)	
	global	local	global	+local	total		$\mathcal{R}$Oxf	$\mathcal{R}$Par
DELG†	3	7	24.0	33.1	57.1	69.0*	4.25	5.35
RRT†	3	7	24.0	33.1	57.1	**3.2**	**2.16**	**2.72**
CVNet	3	**1**	24.0	**1.7**	**25.7**	15.6	27.02	33.55
CVNet^Q	3	**1**	24.0	**1.7**	**25.7**	15.6	6.88	8.52

(e) **Extraction Latency and Memory Footprint.**

Table 4. **Ablation study for CVNet.** mAP measures for each setting. # is the number of samples that are re-ranked.

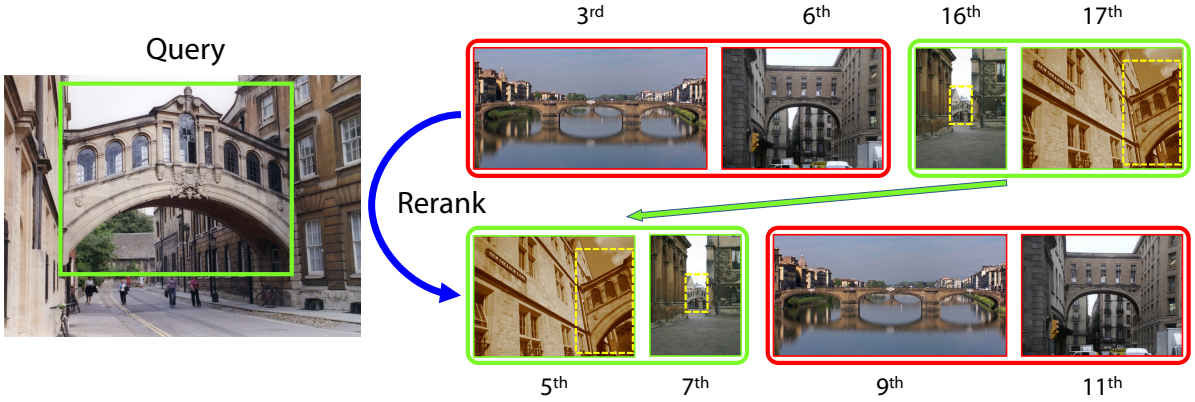

Figure 7. Example qualitative results on $\mathcal{R}$Oxf-Hard+1M with R50-CVNet. The upper row shows the global descriptor matching result and the lower row shows the re-ranking result. Correct/incorrect results are marked with green/red borders, respectively. The query used as an input is generated by cropping only the part bounded by a green square. A dashed yellow line indicates the areas that overlap with the query.

quires considerable improvement. Our future work will aim to achieve improvements in speed and memory while preserving its strong performance.

7. Conclusion

In this study, we propose a novel image retrieval re-ranking network that directly predicts similarity by leveraging dense feature correlation in a convolutional manner. We design the network to construct cross-scale correlations within a single inference, thereby enabling cross-scale matching instead of expensive multi-scale inferences. Considering that re-ranking primarily encounters hard samples during testing, we trained this network by focusing on hard samples. With the aforementioned contributions, we achieve state-of-the-art performance on several benchmarks, demonstrating that dense feature correlation is a powerful cue for image retrieval re-ranking.

Acknowledgements. This work was supported by the Industry Core Technology Development Project, 20005062, Development of Artificial Intelligence Robot Autonomous Navigation Technology for Agile Movement in Crowded Space, funded by the Ministry of Trade, industry & Energy (MOTIE, Republic of Korea).

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5297–5307, 2016. 2

[2] Yannis Avrithis and Giorgos Tolias. Hough pyramid matching: Speeded-up geometry re-ranking for large scale image retrieval. *International Journal of Computer Vision (IJCV)*, 107(1):1–19, 2014. 2

[3] Artem Babenko and Victor Lempitsky. Aggregating local deep features for image retrieval. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 1269–1277, 2015. 2

[4] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural codes for image retrieval. In *Proc. European Conference on Computer Vision (ECCV)*, pages 584–599. Springer, 2014. 2

[5] Axel Barroso-Laguna, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Key. net: Keypoint detection by handcrafted and learned cnn filters. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 5836–5844, 2019. 2

[6] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008. 2

[7] Maxim Berman, Hervé Jégou, Andrea Vedaldi, Iasonas Kokkinos, and Matthijs Douze. Multigrain: a unified image embedding for classes and instances. In *arXiv*, 2019. 2

[8] Bingyi Cao, Andre Araujo, and Jack Sim. Unifying deep local and global features for image search. In *Proc. European Conference on Computer Vision (ECCV)*, pages 726–743. Springer, 2020. 1, 2, 6, 7

[9] Cheng Chang, Guangwei Yu, Chundi Liu, and Maksims Volkovs. Explore-exploit graph traversal for image retrieval. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9423–9431, 2019. 2

[10] Ondrej Chum, James Philbin, Josef Sivic, Michael Isard, and Andrew Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 1–8. IEEE, 2007. 2

[11] 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. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[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. 2

[13] 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 (IJCV)*, 124(2):237–254, 2017. 2, 6

[14] Albert Gordoa, Jose A Rodriguez-Serrano, Florent Perronnin, and Ernest Valveny. Leveraging category-level labels for instance-level image retrieval. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3045–3052. IEEE, 2012. 3

[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. 2, 3, 5

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2, 3

[17] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5901–5910, 2020. 3

[18] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, Teddy Furon, and Ondrej Chum. Efficient diffusion on region manifolds: Recovering small objects with compact cnn representations. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2077–2086, 2017. 2

[19] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proc. European Conference on Computer Vision (ECCV)*, pages 304–317. Springer, 2008. 2

[20] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3304–3311. IEEE, 2010. 2

[21] 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 (TPAMI)*, 34(9):1704–1716, 2011. 2

[22] Shuda Li, Kai Han, Theo W Costain, Henry Howard-Jenkins, and Victor Prisacariu. Correspondence networks with adaptive neighbourhood consensus. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10196–10205, 2020. 2

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

[24] Zixin Luo, Tianwei Shen, Lei Zhou, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. Contextdesc: Local descriptor augmentation with cross-modality context. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2527–2536, 2019. 2

[25] Juhong Min and Minsu Cho. Convolutional hough matching networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2940–2950, 2021. 2, 3

[26] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 2, 4

[27] Anastasiya Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[28] Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Repeatability is not enough: Learning affine regions via discriminability. In *Proc. European Conference on Computer Vision (ECCV)*, pages 284–300, 2018. 2

[29] Tony Ng, Vassileios Balntas, Yurun Tian, and Krystian Mikolajczyk. Solar: second-order loss and attention for image retrieval. In *Proc. European Conference on Computer Vision (ECCV)*, pages 253–270. Springer, 2020. 6

[30] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 3456–3465, 2017. 1, 2, 5, 6

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems (NeurIPS)*, 32:8026–8037, 2019. 5

[32] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2007. 1, 2, 6

[33] 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. 2, 6

[34] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5706–5715, 2018. 6

[35] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *Proc. European Conference on Computer Vision (ECCV)*, pages 3–20. Springer, 2016. 2, 3

[36] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(7):1655–1668, 2018. 2, 6

[37] Jerome Revaud, Jon Almazán, Rafael S Rezende, and Cesar Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 5107–5116, 2019. 6

[38] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 5

[40] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4938–4947, 2020. 6

[41] 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. 1, 2, 6

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. International Conference on Learning Representations (ICLR)*, 2015. 3

[43] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017. 2, 5

[44] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, volume 3, pages 1470–1470. IEEE Computer Society, 2003. 2

[45] 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, 2, 6, 7

[46] 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. 1, 2, 6, 7

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

[48] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular object retrieval with integral max-pooling of cnn activations. In *Proc. International Conference on Learning Representations (ICLR)*, 2015. 2

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017. 2

[50] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2575–2584, 2020. 5, 6, 7

[51] Chang Xu, Yangxi Li, Chao Zhou, and Chao Xu. Learning to rerank images with enhanced spatial verification. In *Proc. IEEE International Conference on Image Processing (ICIP)*, pages 1933–1936. IEEE, 2012. 2

[52] Gengshan Yang and Deva Ramanan. Volumetric correspondence networks for optical flow. *Advances in Neural Information Processing Systems (NeurIPS)*, 32:794–805, 2019. 2

[53] 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 *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 11772–11781, 2021. 1, 2, 6

[54] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *Proc. European Conference on Computer Vision (ECCV)*, pages 467–483. Springer, 2016. 2

[55] Shuhei Yokoo, Kohei Ozaki, Edgar Simo-Serra, and Satoshi Iizuka. Two-stage discriminative re-ranking for large-scale landmark retrieval. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1012–1013, 2020. 5, 7