

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 ROxford-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

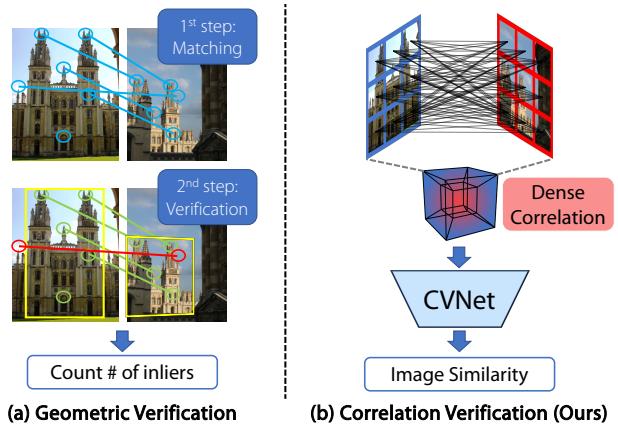


Figure 1. Novel image retrieval re-ranking method named correlation verification that directly predicts image similarity by leveraging dense feature correlation in a convolutional manner.

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.

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: ROxford (+1M), RParis (+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 i th *ResBlock*.

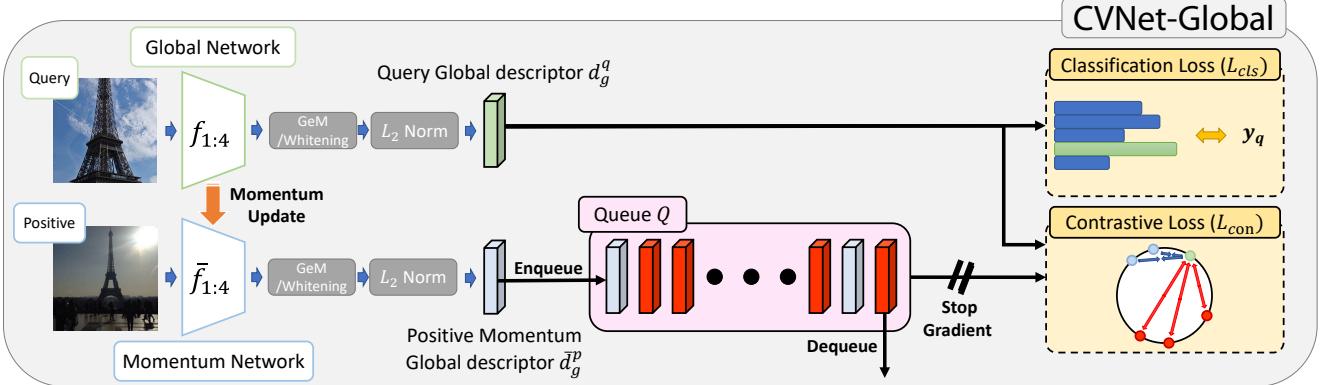


Figure 2. Illustration of the proposed Global backbone network (CVNet-Global) and its training objective. The network has two objectives: classification loss and contrastive loss. To utilize several samples without a computational burden in contrastive learning, momentum network and queue structure are 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 $\mathbf{Q} \in \{\bar{d}_g^i\}_{i=1}^K$, 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 d_g^q . With d_g^q , CurricularFace [17]-marginized classification loss \mathcal{L}_{cls} is computed as

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

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 i th class y_i is identical to y_g . \mathcal{C} is a function that adds a CurricularFace margin to cosine similarity with its margin term m .

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{d}_g^p . The descriptor \bar{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{d}_g^p . Thus we use the CurricularFace-margined momentum contrastive loss \mathcal{L}_{con} :

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

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 $s_t^{q,k} \in \mathbb{R}^1$ 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

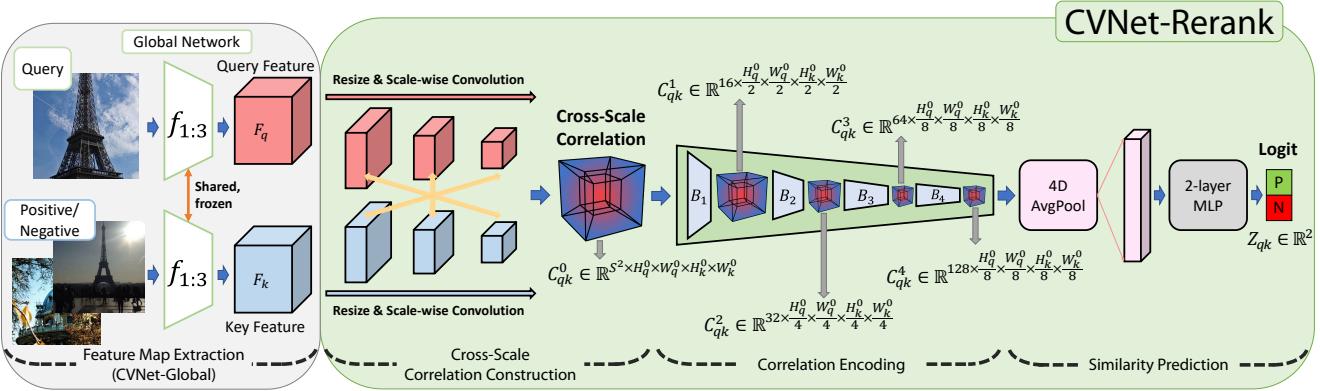


Figure 3. Illustration of the proposed Re-ranking network (CVNet-Rerank). The proposed network takes pair of feature maps extracted from the trained CVNet-Global model as input, constructs a cross-scale feature correlation, and gradually compresses it to image similarity of a pair with deeply stacked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

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

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

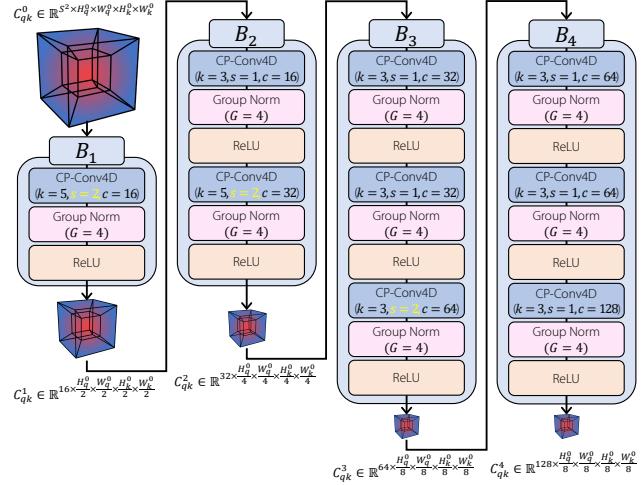


Figure 4. The detailed structure of the proposed 4D correlation Encoder. The proposed encoder structure gradually compresses the cross-scale correlation into a fine-grained correlation cue.

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) :

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

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:

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



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.

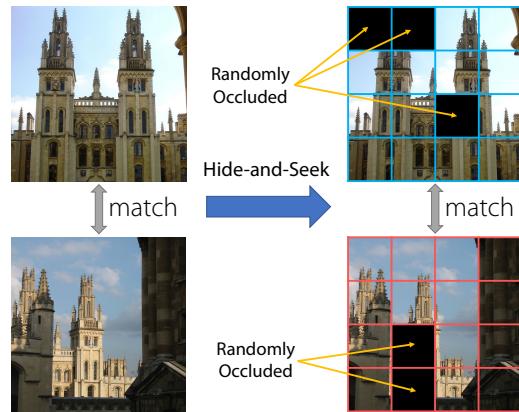


Figure 6. With Hide-and-Seek, the re-ranking network can effectively learn hard-matching cases by randomly hiding parts of matching pairs to give images an occlusion-like effect.

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	
	$\mathcal{R}\text{Oxf}$	+1M	$\mathcal{R}\text{Par}$	+1M	$\mathcal{R}\text{Oxf}$	+1M	$\mathcal{R}\text{Par}$	+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 † [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 $\mathcal{R}\text{Oxf}/\mathcal{R}\text{Par}$ 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 $\mathcal{R}\text{Oxf}/\mathcal{R}\text{Par}$ 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 $\mathcal{R}\text{Oxford5k}$ [32, 34] (referred to as $\mathcal{R}\text{Oxf}$) and $\mathcal{R}\text{Paris6k}$ [33, 34] (referred to as $\mathcal{R}\text{Par}$) datasets. Both datasets comprise 70 queries and 4933 and 6322 database images, respectively. In addition, an $\mathcal{R}\text{1M}$ 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 $\mathcal{R}\text{Oxf}$ and $\mathcal{R}\text{Par}$, 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 ($\mathcal{R}\text{Oxf}$), 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- $\mathcal{R}\text{Oxf}$ -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 **bold-faced** 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

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

reduces the memory footprint by 1/4, it hardly diminishes the overall performance.

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
200		✓	86.1	77.6	89.4	79.9	72.8	61.1	78.6	63.9
	✓		81.3	72.6	88.7	78.9	62.5	50.2	76.5	60.2
400			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
			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).

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

\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

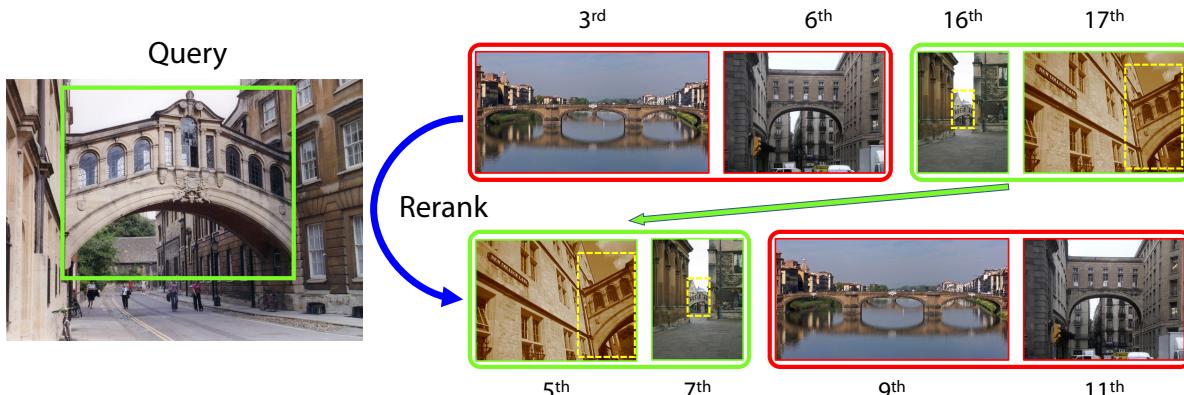
(c) Loss Comparison of CVNet-Global.

#	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
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.

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 sam-

ples 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 hand-crafted 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 Gordo, 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, Shaolin 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](#)

Supplementary Material: 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

S1. Data Selection and Sampling Process

Overlapped positive selection. In this study, we use the ‘clean’ subset [18] of Google Landmarks dataset v2 (1.58M images from 81k landmarks) [16] as a training set. This dataset has large intra-class variability and includes multiple viewpoints, such as indoor and outdoor views of landmarks. Therefore, when sampling the same-class image pair from this dataset, we cannot guarantee an overlap between the two images, and non-overlapping query-positive pairs can interfere with learning image matching. To avoid the non-overlapping case, we select overlapped pairs for each class in advance with the help of the DELF [9] local feature. The overall process is similar to the data cleaning process of [18]: The primary difference is that [18] aims to remove outlier data from the dataset, whereas we aim to select same-class pairs that actually overlap. To select an overlapped pair, for every dataset sample x_i , we first select up to ten of the nearest neighbors that are assigned to the same class as x_i with a global descriptor extracted from R50-CVNet-Global. After the nearest neighbors are selected, spatial verification using RANSAC with a pre-trained DELF feature is performed on the nearest neighbors selected for each sample. Subsequently, we select the pair with 30 or more inlier matches as an overlapped pair. Furthermore, only classes with more than 10 samples belonging to overlapped pairs are used for training. Finally, we select 1M images from 31k landmarks of the GLDv2-clean dataset and use this subset as a training set for CVNet-Rerank. Although this selection process is quite expensive because of the use of RANSAC, it only needs to be performed once.

Sampling process. CVNet-Rerank is trained for 200 epochs (6.3M steps) for all selected classes. For every epoch, we construct tuples of query, positive, and negative samples for each class. The query image is randomly sampled from each class, a positive image is randomly chosen from among the overlapped positives of the query, and a negative image is sampled from random or hard-negative

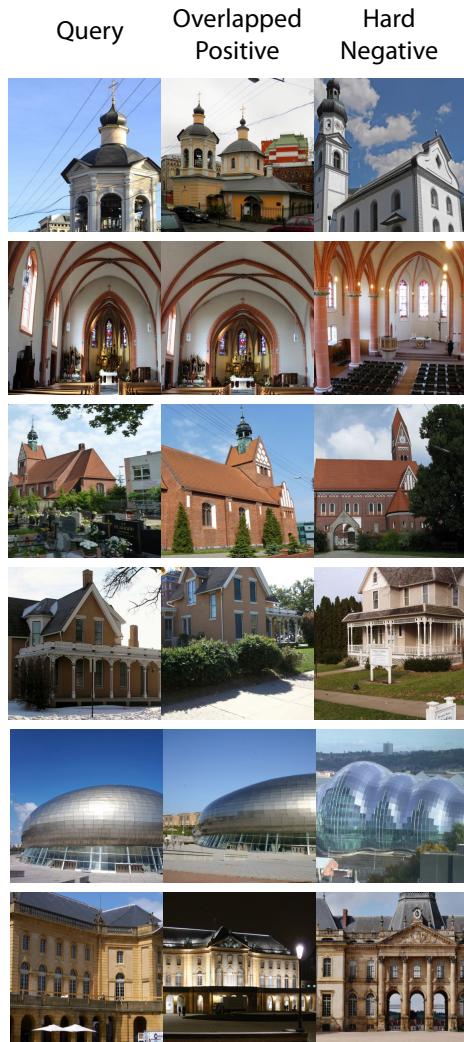
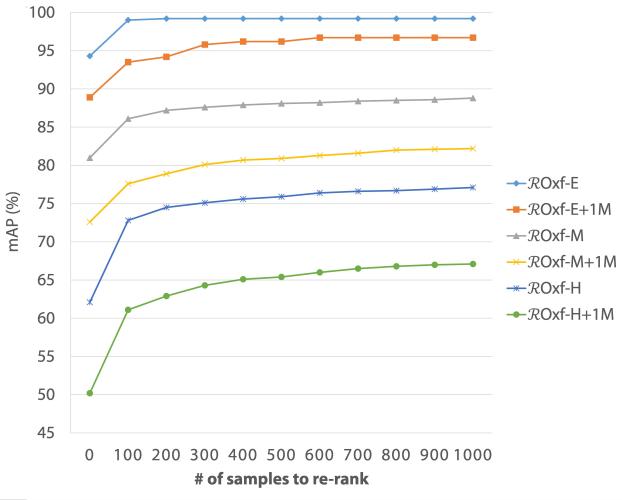
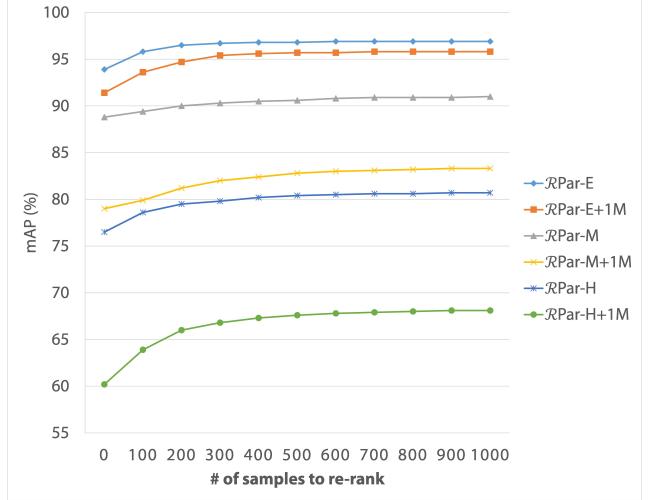


Figure S1. Example of query, overlapped positive, and hard negative samples sampled from the selected subset of the GLDv2-clean dataset. Our proposed re-ranking network learns better discrimination ability by learning the cue for equivalence from a overlapped positive and the cue for the difference from a hard negative.

*Corresponding author.



(a) ROxford5k and its +1M Experiments.



(b) RParis6k and its +1M Experiments.

Figure S2. Analysis about number of samples to re-rank.

rerank	r_H	Medium			Hard				
		ROxf +1M	RPar +1M	ROxf +1M	RPar +1M	ROxf +1M	RPar +1M		
0	0.0	81.0	72.6	88.8	79.0	62.1	50.2	76.5	60.2
100	0.0	81.6	72.8	88.8	79.0	62.6	50.2	76.6	60.3
	1.0	85.5	77.1	89.3	80.0	70.8	59.5	77.5	63.7
	0.5	85.5	77.3	89.2	79.8	71.5	60.5	77.9	63.4
	0.2-1.0	86.1	77.6	89.4	79.9	72.8	61.1	78.6	63.9
200	0.0	81.5	72.7	88.8	79.0	62.6	50.1	76.7	60.4
	1.0	86.2	78.2	89.5	81.1	71.5	60.7	76.5	65.3
	0.5	86.4	78.4	89.7	80.8	73.1	62.0	78.7	65.3
	0.2-1.0	87.2	78.9	90.0	81.2	74.5	62.9	79.5	66.0
400	0.0	81.4	72.6	88.9	79.1	62.6	50.1	77.1	60.6
	1.0	86.2	79.5	88.5	81.8	71.0	62.0	74.0	65.5
	0.5	86.9	79.8	90.3	82.0	74.0	63.9	79.5	66.7
	0.2-1.0	87.9	80.7	90.5	82.4	75.6	65.1	80.2	67.3

Table S1. Hard-Negative Sampling Ratio.

samples according to the hard negative sampling ratio r_{neg} . Fig. S1 shows examples of our sampling results. By learning with well-constructed training pairs, the network can achieve improved discriminating ability.

S2. Additional Ablation Studies and Analysis

S2.1. Curriculum Learning

Learning focused on hard samples can improve the robustness of the network in hard situations. However, this could lead to a loss of generality. Accordingly, we apply curriculum learning to focus on hard samples without losing generality. In this subsection, we show that the proposed network performs re-ranking well regardless of the matching difficulty with the help of curriculum learning. Furthermore, we show a more detailed analysis of curriculum learning.

Generality of learning (Fig. S2). By gradually increasing the number of samples to be re-ranked, we can verify whether the network distinguishes hard samples well

rerank	$phas$	Medium			Hard				
		ROxf +1M	RPar +1M	ROxf +1M	RPar +1M	ROxf +1M	RPar +1M		
0	0.0	81.0	72.6	88.8	79.0	62.1	50.2	76.5	60.2
100	0.2	85.8	77.5	89.3	79.9	71.6	60.5	78.1	63.7
	0.0-0.2	86.1	77.6	89.4	79.9	72.8	61.1	78.6	63.9
	0.0	86.9	78.7	89.7	81.0	73.4	62.1	78.6	65.6
	0.2	87.1	78.3	89.7	81.1	74.0	61.8	78.7	65.7
200	0.0-0.2	87.2	78.9	90.0	81.2	74.5	62.9	79.5	66.0
	0.0	87.5	80.3	89.9	82.0	74.2	64.3	78.9	66.4
	0.2	87.8	80.1	90.1	82.2	75.1	64.0	79.3	66.8
	0.0-0.2	87.9	80.7	90.5	82.4	75.6	65.1	80.2	67.3

Table S2. Hide-and-Seek Probability.

while retaining generality for normal samples. As shown in Fig. S2, the proposed re-ranking network dramatically improves performance when it is applied to top ranks where many hard samples exist. Even if the re-ranking targets are expanded to easier samples, our proposed re-ranking model continues to exhibit improved performance without losing generality.

Hard negative mining (Tab. S1). To prove the effectiveness of hard negative mining applied simultaneously with the curriculum approach, we conduct experiments by varying the hard-negative sampling ratio r_H . The results are presented in Tab. S1. When the network learns using randomly sampled negatives ($r_H = 0$), global retrieval results do not improve when re-ranking. This indicates that learning to discriminate hard samples using only random negative is difficult. Accordingly, when sampling hard negatives with a fixed ratio ($r_H = 1.0, 0.5$), the network exhibits a significantly improved performance. Moreover, when a hard-negative ratio is set through the curriculum manner ($r_H = 0.2-1.0$), the proposed re-ranking network exhibits its best performance. This proves that hard negatives are a critical key to re-ranking learning, and hard negative mining

#	C'_l	Medium				Hard			
		$\mathcal{R}\text{Oxf}$	+1M	$\mathcal{R}\text{Par}$	+1M	$\mathcal{R}\text{Oxf}$	+1M	$\mathcal{R}\text{Par}$	+1M
0	81.0	72.6	88.8	79.0	62.1	50.2	76.5	60.2	
	16	85.4	76.3	89.2	79.7	70.9	57.9	77.5	62.5
	32	85.5	76.5	89.2	79.8	71.7	59.7	77.6	63.1
	64	85.4	76.8	89.3	79.9	71.6	60.2	77.6	63.3
100	128	85.5	76.9	89.3	79.9	71.4	60.0	77.9	63.6
	256	86.1	77.6	89.4	79.9	72.8	61.1	78.6	63.9
	512	85.5	76.9	89.3	79.9	71.3	59.7	77.8	63.7
	1024	85.7	77.3	89.4	80.0	71.6	59.7	78.1	63.7
	16	86.2	77.2	89.6	80.5	71.9	58.8	78.1	63.9
	32	86.5	77.6	89.6	80.7	73.2	61.1	78.1	64.5
	64	86.2	77.9	89.7	80.9	72.9	61.3	77.9	64.9
200	128	86.4	78.0	89.9	81.2	72.8	61.2	78.6	65.5
	256	87.2	78.9	90.0	81.2	74.5	62.9	79.5	66.0
	512	86.3	77.9	89.8	81.1	72.6	61.0	78.4	65.6
	1024	86.5	78.6	89.9	81.1	72.6	61.2	79.1	65.5
	16	86.5	78.4	89.9	81.2	72.3	60.2	78.5	64.5
	32	87.0	79.1	89.7	81.5	74.1	63.0	78.2	65.1
	64	86.7	79.3	89.8	81.7	73.6	63.3	78.1	65.6
400	128	87.0	79.6	90.1	82.1	73.5	63.2	78.9	66.3
	256	87.9	80.7	90.5	82.4	75.6	65.1	80.2	67.3
	512	86.6	79.3	90.0	82.0	73.1	62.8	78.5	66.3
	1024	87.0	80.1	90.3	82.1	73.4	63.0	79.6	66.6

Table S3. Channel Compression.

is even more effective when used with curriculum learning.

Hide-and-Seek (Tab. S2). Similarly, to prove the effectiveness of the Hide-and-Seek [13] augmentation, we conduct experiments by varying the Hide-and-Seek probability p_{has} . Tab. S2 also shows that Hide-and-Seek is an appropriate strategy to help re-ranking learning and that it can be even more effective when used with curriculum learning.

S2.2. Memory Footprint Reduction

Despite having significant potential, the proposed re-ranking method possesses a large memory owing to its dense nature. In this subsection, we present several effective solutions for reducing the memory footprint of the proposed re-ranking model.

Channel compression (Tab. S3). We pre-extract and store a multi-scale feature pyramid for every database sample for online re-ranking, which is where memory consumption primarily occurs. To reduce the memory footprint of the proposed model, we compress the channel of the feature map C_l to C'_l using a 3×3 convolution layer in the process of constructing the multi-scale feature pyramid. Here, we conduct experiments by varying the compressed channel dimension C'_l , to find a balance between memory footprint and re-ranking performance. The results are presented in Tab. S3. When the C'_l is 256, the proposed re-ranking model exhibited its best performance; therefore, we finally selected C'_l as 256 in our study. However, on systems where memory management is more important, choosing a smaller

#	type	Medium				Hard			
		$\mathcal{R}\text{Oxf}$	+1M	$\mathcal{R}\text{Par}$	+1M	$\mathcal{R}\text{Oxf}$	+1M	$\mathcal{R}\text{Par}$	+1M
0		81.0	72.6	88.8	79.0	62.1	50.2	76.5	60.2
	float32	86.1	77.6	89.4	79.9	72.8	61.1	78.6	63.9
100	int8	86.1	77.6	89.4	79.9	72.7	61.1	78.6	63.9
	int4	86.0	77.3	89.3	79.8	72.5	60.5	78.0	63.5
	float32	87.2	78.9	90.0	81.2	74.5	62.9	79.5	66.0
200	int8	87.2	78.9	90.0	81.2	74.5	62.8	79.5	66.0
	int4	86.9	78.6	89.7	81.0	73.8	62.1	78.7	65.3
	float32	87.9	80.7	90.5	82.4	75.6	65.1	80.2	67.3
400	int8	87.9	80.6	90.5	82.4	75.5	65.1	80.2	67.3
	int4	87.4	80.1	90.1	82.0	74.6	63.8	79.3	66.3

Table S4. Feature Quantization.

#	layer	Medium				Hard			
		$\mathcal{R}\text{Oxf}$	+1M	$\mathcal{R}\text{Par}$	+1M	$\mathcal{R}\text{Oxf}$	+1M	$\mathcal{R}\text{Par}$	+1M
0		81.0	72.6	88.8	79.0	62.1	50.2	76.5	60.2
100	f_3	86.1	77.6	89.4	79.9	72.8	61.1	78.6	63.9
	f_4	85.4	76.0	89.2	79.8	69.0	56.2	76.7	63.0
	fuse	85.2	76.8	89.3	79.9	69.6	57.6	77.0	63.2
200	f_3	87.2	78.9	90.0	81.2	74.5	62.9	79.5	66.0
	f_4	85.8	76.7	89.2	80.7	69.7	57.0	75.5	63.8
	fuse	85.4	77.5	89.4	80.9	69.8	58.2	75.8	64.0
400	f_3	87.9	80.7	90.5	82.4	75.6	65.1	80.2	67.3
	f_4	86.0	77.7	88.2	80.9	69.8	58.0	73.7	63.2
	fuse	85.1	78.3	88.2	80.9	68.9	58.3	74.0	63.3

Table S5. Feature Extraction Layer Selection.

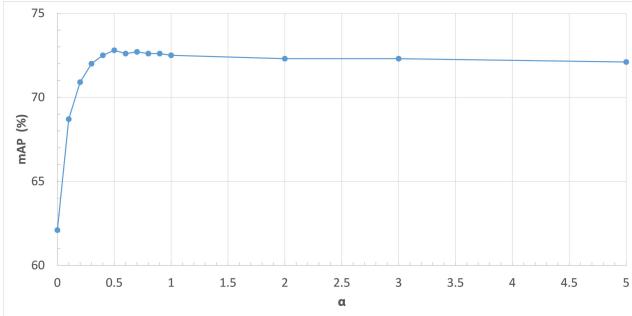
dimension, such as 16 ($\frac{1}{16}$ of our model’s memory footprint) or 32 ($\frac{1}{8}$ of our model’s memory footprint) can be a good option. This quantization reduces the memory footprint even further, albeit at the cost of a marginally reduced performance.

Quantization (Tab. S4). To reduce the memory burden, we measured the re-ranking performance while taking the correlation of quantized features as an input. The results are presented in Tab. S4. Similar to the case of channel compression, feature quantization also reduces the memory footprint at the risk of marginal performance degradation.

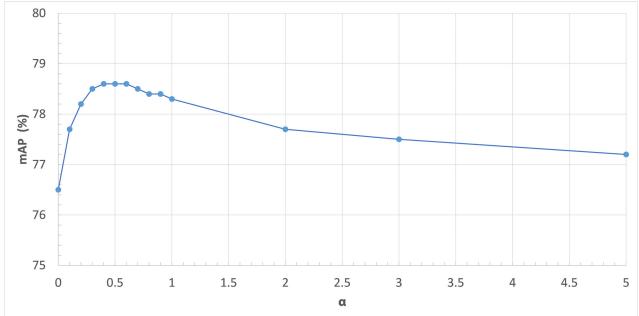
S2.3. Model Design and Parameter Selection

In this subsection, we present several analyses of the design of the re-ranking model and its parameter selection.

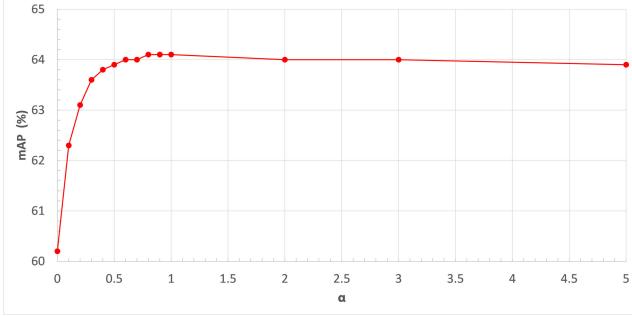
Feature extraction layer selection (Tab. S5). First, we analyze CVNet-Global to determine which of its stages is more suited for use as an input for the re-ranking network. The results are presented in Tab. S5. f_i denotes the i th *Res-Block*. When receiving an output of f_4 as an input, the stride and kernel size in the first block are reduced by 1 and 3, respectively; therefore, the output resolution is identical to that when an output of f_3 is received as an input. In the “fuse” case, both the output feature maps of f_3 and f_4 are received as input. In this case, the outputs of f_3 and f_4 pass through the first two convolutional blocks separately and



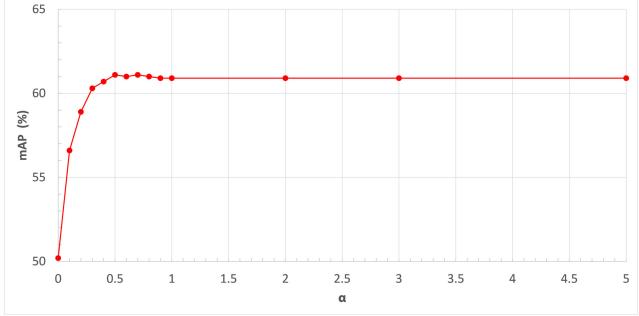
(a) ROxford5k-Hard Experiments.



(b) RParis6k-Hard Experiments.



(c) ROxford5k-Hard+1M Experiments.



(d) RParis6k-Hard+1M Experiments.

Figure S3. Experiments about the score fusion weight α . α is tuned in ROxf-Hard (Fig. S3a)/RPar-Hard (Fig. S3b) and fixed for ROxf-Hard+1M (Fig. S3c)/RPar-Hard+1M (Fig. S3d). We finally set an α to 0.5.

#	Scale	Medium			Hard				
		$\frac{1}{1}$	$\frac{1}{2}$	$\frac{1}{\sqrt{2}}$	ROxf +1M	RPar +1M	ROxf +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
	✓ ✓	85.8	77.1	89.3	80.0	71.5	59.8	78.3	63.9
	✓ ✓ ✓	85.7	77.3	89.3	79.9	71.1	59.9	78.0	63.4
	✓ ✓ ✓ ✓	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
	✓ ✓	86.6	78.2	90.0	81.2	72.8	61.3	79.0	66.0
	✓ ✓ ✓	86.5	78.5	89.9	81.0	72.2	61.1	78.7	65.3
	✓ ✓ ✓ ✓	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
	✓ ✓	87.1	79.9	90.3	82.4	73.6	63.4	79.3	67.2
	✓ ✓ ✓	87.0	80.2	90.3	82.0	72.9	63.1	79.0	66.4
	✓ ✓ ✓ ✓	87.9	80.7	90.5	82.4	75.6	65.1	80.2	67.3

Table S6. Scale Selection.

merged, and finally pass through the remaining blocks. As in many studies [1, 9, 17] utilizing local information, using the output of f_3 as an input results in the best performance; thus, we select the feature map from f_3 as the input of the re-ranking network.

Scale selection (Tab. S6). We conduct experiments with a selection of scales to construct a multi-scale feature pyramid. Note that a high-scale feature can be helpful in terms of performance. However, considering the limitation of time and memory, we only scale the feature to a lower scale. The results show that constructing a cross-scale correlation using several scales has a clear performance advantage over

#	kernel	Medium			Hard				
		ROxf +1M	RPar +1M	ROxf +1M	RPar +1M	ROxf +1M	RPar +1M		
0		81.0	72.6	88.8	79.0	62.1	50.2	76.5	60.2
100	asymmetric	86.1	77.6	89.4	79.9	72.8	61.1	78.6	63.9
	symmetric	85.2	76.8	89.3	79.9	69.6	57.6	77.0	63.2
200	asymmetric	87.2	78.9	90.0	81.2	74.5	62.9	79.5	66.0
	symmetric	85.4	77.5	89.4	80.9	69.8	58.2	75.8	64.0
400	asymmetric	87.9	80.7	90.5	82.4	75.6	65.1	80.2	67.3
	symmetric	85.1	78.3	88.2	80.9	68.9	58.3	74.0	63.3

Table S7. Kernel Symmetrization.

the single-scale feature correlation method. Based on the experimental results, we finally select $S = 3$ scales.

Symmetric kernel (Tab. S7). Image similarity is essentially permutation-invariant, except in special cases. When we train a 4D convolutional network to predict image similarity, we can induce the network to be permutation-invariant in several ways. For instance, we can set the loss function to ensure that the output does not vary regardless of the input order. Another method is to make the 4D convolution kernel symmetrical. We experiment with the latter case as shown in Tab. S7. However, forcing the kernel to be symmetric did not yield good performance. Therefore, in this study, we softly induce permutation-invariant properties in the re-ranking network using loss symmetrization.

Score fusion weight (Fig. S3). To simultaneously verify the global and local relationships between two images, we re-rank the retrieval results based on the combined score $s_g + \alpha s_r$, where s_g is the cosine similarity of the global de-

#	layer	Medium				Hard			
		$\mathcal{R}\text{Oxf}$	+1M	$\mathcal{R}\text{Par}$	+1M	$\mathcal{R}\text{Oxf}$	+1M	$\mathcal{R}\text{Par}$	+1M
0	Global	81.0	72.6	88.8	79.0	62.1	50.2	76.5	60.2
0	αQE	85.4	77.5	90.7	83.5	67.5	57.8	79.8	66.9
100	CV	86.1	77.6	89.4	79.9	72.8	61.1	78.6	63.9
200	CV	87.2	78.9	90.0	81.2	74.5	62.9	79.5	66.0
400	CV	87.9	80.7	90.5	82.4	75.6	65.1	80.2	67.3
0	αQE	85.4	77.5	90.7	83.5	67.5	57.8	79.8	66.9
100	$\alpha\text{QE} + \text{CV}$	88.0	80.5	90.9	84.1	74.6	65.2	80.9	70.0
200	$\alpha\text{QE} + \text{CV}$	88.8	82.1	91.2	84.9	75.9	67.4	81.6	71.5
400	$\alpha\text{QE} + \text{CV}$	89.3	82.8	91.6	85.3	77.1	68.6	82.2	70.7

Table S8. Comparison with alphaQE.

scriptors, s_r is the output score of the re-ranking network, and α is the given weight for the re-ranking network output score s_r . Parameter α is tuned in $\mathcal{R}\text{Oxf}/\mathcal{R}\text{Par}$ and fixed for a large-scale experiment and GLDv2-retrieval test, as in previous studies [1, 8, 12, 15]. Fig. S3a and Fig. S3b shows $\mathcal{R}\text{Oxf-Hard}/\mathcal{R}\text{Par-Hard}$ performances according to score fusion weight α . In these results, the re-rank score significantly improves the retrieval performance even if an extremely small re-rank score is added to the global descriptor matching score. Moreover, the best performance corresponded to an α value of 0.5. Based on these experimental results, we set $\alpha = 0.5$ for the re-ranking process.

S2.4. Comparison with Query Expansion

Comparison with αQE (Tab. S8). This study focuses on improving the image matching ability for single pairs. Therefore, we have not considered certain re-ranking methods such as diffusion [2, 5] or query expansion [3, 11], which require additional expenses to traverse the entire database mentioned in the main body of this paper. Although we do not consider them because of their different scopes, in this subsection we show that these re-ranking methods and the proposed re-ranking method can be harmoniously fused when they are used. Specifically, we compared and fused CV with one of the representative query expansion methods: α -weighted query expansion (αQE).

In contrast to geometric verification (GV) or our proposed correlation verification (CV), which evaluates the similarity between *two images*, the query expansion aggregates the query itself and its top-ranked neighbors *across the dataset* and creates an aggregated query to perform re-ranking. In the αQE method, aggregation is performed with weighted averaging, and the weight of the i th ranked image is given by $(\mathbf{d}_q \cdot \mathbf{d}_i)^{\alpha\text{QE}}$, where \mathbf{d}_q is the global descriptor of the query image and \mathbf{d}_i is the global descriptor of the i th ranked image for the query. Finally, the aggregated query descriptor \mathbf{d}'_q is computed as follows:

$$\mathbf{d}'_q = \frac{\mathbf{d}_q + \sum_{i=1}^n ((\mathbf{d}_q \cdot \mathbf{d}_i)^{\alpha\text{QE}} \cdot \mathbf{d}_i)}{1 + \sum_{i=1}^n (\mathbf{d}_q \cdot \mathbf{d}_i)^{\alpha\text{QE}}}, \quad (\text{S1})$$

where n is the number to aggregates, and αQE is a pa-

K	Medium				Hard			
	$\mathcal{R}\text{Oxf}$	+1M	$\mathcal{R}\text{Par}$	+1M	$\mathcal{R}\text{Oxf}$	+1M	$\mathcal{R}\text{Par}$	+1M
576	77.5	70.0	89.8	78.0	55.7	44.6	77.9	58.9
4608	78.5	71.3	89.8	78.7	58.1	46.4	78.3	59.3
73728	81.0	72.6	88.8	79.0	62.1	50.2	76.5	60.2

Table S9. Queue Size of Momentum Contrastive Loss.

Loss	Medium				Hard			
	$\mathcal{R}\text{Oxf}$	+1M	$\mathcal{R}\text{Par}$	+1M	$\mathcal{R}\text{Oxf}$	+1M	$\mathcal{R}\text{Par}$	+1M
SupCon [6] (Eq. (S2))	79.9	72.1	89.4	78.9	59.1	48.6	77.4	59.3
Ours (Eq. (S3))	81.0	72.6	88.8	79.0	62.1	50.2	76.5	60.2

Table S10. Comparison with SupCon Loss.

rameter that amplifies or reduces the weight. n and αQE are tuned in $\mathcal{R}\text{Oxf}/\mathcal{R}\text{Par}$ over the ranges: $n \in [1, 20]$ and $\alpha\text{QE} \in [0.1, 2.0]$ and we finally set them to 5 and 2.0, respectively. Tab. S8 shows the re-ranking results using αQE , the re-ranking results using CV, and the re-ranking results using αQE and CV sequentially. For all settings, CV exhibits performance that is superior to the αQE method, and even more superior when fused with the αQE method.

S2.5. Momentum Contrastive Loss Analysis

Queue size (Tab. S9). Our global backbone network, CVNet-Global, constructs a queue to store and leverage numerous samples for contrastive learning. Because queue size is one of the crucial factors that is directly related to the number of contrastive samples, we conduct experiments by varying the queue size K . Tab. S9 shows performances for different queue sizes. Overall, our global model benefits from a large K value. A large queue size implies that several contrastive samples can be utilized, which can lead the global model to learn a more generalized representation.

Differences in SupCon loss (Tab. S10). Our momentum contrastive loss is similar to SupCon [6] loss. Similar to the SupCon loss, it performs contrast learning with multiple positives using labels. The primary difference between these losses is that the SupCon loss assumes a relatively constant number of positives. However, a large difference exists in the number of positives for each sample because of the class imbalance data and queue structure. In SupCon loss (Eq. (S2)), because all query-positive cosine similarities are included in the denominator, the scale of the loss is significantly affected by the number of positives:

$$\mathcal{L}_s = \frac{-1}{|P(q)|} \sum_{p \in P(q)} \log \frac{\exp(\bar{\mathcal{C}}(\mathbf{d}_q^q \cdot \bar{\mathbf{d}}_g^p, 1)/\tau)}{\sum_{i \in P(q) \cup N(q)} \exp(\bar{\mathcal{C}}(\mathbf{d}_q^q \cdot \bar{\mathbf{d}}_g^i, \mathbb{1}_q^i)/\tau)}. \quad (\text{S2})$$

To solve this scale problem, we design our contrastive loss (Eq. (S3)) similar to the SupCon loss \mathcal{L}_s . However, only the target positive p is included in the denominator.

model	Medium			Hard				
(Rerank top-100)	$\mathcal{R}\text{Oxf}$	+1M	$\mathcal{R}\text{Par}$	+1M	$\mathcal{R}\text{Oxf}$	+1M	$\mathcal{R}\text{Par}$	+1M
R50-DELG [†]	71.1	60.4	86.9	70.9	47.0	32.0	73.6	48.1
+ CVNet-Rerank	78.7	67.7	87.9	72.3	63.0	46.1	76.8	52.5
R50-DOLG [†]	79.0	70.0	88.3	76.2	57.5	43.2	75.0	55.4
+ CVNet-Rerank	83.7	74.9	89.0	77.2	69.1	55.7	77.1	59.0

Table S11. Performance with Different Backbones.

model (R50)	Medium			Hard				
(Rerank top-100)	$\mathcal{R}\text{Oxf}$	+1M	$\mathcal{R}\text{Par}$	+1M	$\mathcal{R}\text{Oxf}$	+1M	$\mathcal{R}\text{Par}$	+1M
CVNet-Global	81.0	72.6	88.8	79.0	62.1	50.2	76.5	60.2
+ CorrNet	81.3	72.7	88.8	79.0	62.3	50.3	76.5	60.2
+ HNM	84.6	75.9	89.0	79.3	69.3	57.0	76.9	61.1
+ CSC	85.8	77.5	89.3	79.9	71.6	60.5	78.1	63.7
+ HaS	86.1	77.6	89.4	79.9	72.8	61.1	78.6	63.9

Table S12. Module Ablation Study.

$$\mathcal{L}_{con} = \frac{-1}{|P(q)|} \sum_{p \in P(q)} \log \frac{\exp(\bar{\mathcal{C}}(\mathbf{d}_g^q \cdot \bar{\mathbf{d}}_g^p, 1) / \tau)}{\sum_{i \in [p] \cup N(q)} \exp(\bar{\mathcal{C}}(\mathbf{d}_g^q \cdot \bar{\mathbf{d}}_g^i, \mathbb{1}_q^i) / \tau)}. \quad (\text{S3})$$

Tab. S10 shows the results of training CVNet-Global using each of the two losses. When using the proposed contrastive loss \mathcal{L}_{con} , the results are more stable than when using SupCon loss L_s .

S2.6. Performance with Different Backbones

Tab. S11 shows the results when CVNet-Rerank is combined with the standard global models, DELG and DOLG (both reproduced[†]). Both models are trained using the settings of the original paper except for setting the maximum number of epochs to 25 epochs. Afterward, the proposed CVNet-Rerank is trained with each global backbone. As shown in Tab. S11, CVNet-Rerank also works well when combined with other global backbones.

S2.7. Module Ablation Study

Tab. S12 shows the results when the components of CVNet-Rerank are added to CVNet-Global one by one. From the Tab. S12, we can see that when the Correlation encoding Network (CorrNet) is added, the accuracy is slightly improved but when Hard Negative Mining (HNM) is applied, the accuracy is significantly improved. From the observation, we can see that the correlation encoding network and hard negative mining is a good combination but the network is quite hard to train if the hard negative mining is not used. Based on our experience, when we train the network on random negatives without hard negative mining, the training is dominated by the very rare high correlations between the query and the random negatives degrading the accuracy. To prevent the degradation, we have to train the network using hard negative mining. In summary, we can say that the correlation encoding network and hard negative mining is the core component of the CVNet-Rerank, and

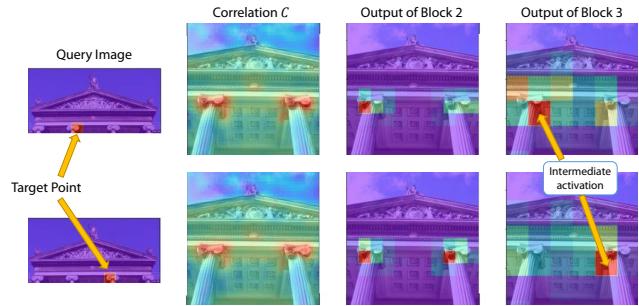


Figure S4. **Intermediate Feature Visualization.** Our network naturally learns the correct geometric relationship of dense matching and pays attention to the correct position by compressing the surrounding matching information from the 4D correlation.

the combination significantly improves the accuracy. Cross-Scale Correlation (CSC) and Hide-and-Seek (HaS) are the optional choices that can incrementally improve the accuracy of the re-ranking network.

S3. Intermediate Feature Visualization

We visualize the intermediate features of our re-ranking model to see how the network interprets and compresses the correlation. To visualize the intermediate 4D features, we select one target point from the query side and visualize the magnitude of the corresponding feature parts on the key side. The visualized results are presented in Fig. S4. In Fig. S4, we observe that the model focus on the correct position by compressing the surrounding matching information from the 4D correlation. As shown in the results, the network naturally learns the geometric pattern of dense matching without any predefined geometric model (e.g. Affine model). Additional intermediate feature visualizations are presented in Fig. S5.

S4. Reproducing Details

For a fair comparison with other re-ranking methods, we conduct experiments by reproducing other re-ranking methods based on the global backbone network. We reproduce two re-ranking methods: geometric verification (GV) and Reranking Transformer [14]. Because both methods are based on the local features of DELG, we attach the local branch of DELG [1] to our global backbone (R50-CVNet-Global) to learn the local features of DELG. All local-feature-related settings are identical to those in the DELG [1]. During testing, we extract a maximum of 1000 local features (500 for RRT) and use them for the re-ranking process.

Geometric Verification (GV). We reproduce the GV based on the DELG. Official code of DELG uses RANSAC [4], which belongs to the scikit-learn [10] package; however, we could not improve the re-ranking performance with

this version. Finally, we implement RANSAC using pydengensac [7], which exhibits performance superior to that of scikit-learn. Additionally, as mentioned in DELG [1] paper, we set a minimum number of inliers to improve re-ranking performance. We tune the minimum number of inliers over the range: [10,300], and finally set it to be 100.

Ranking Transformers (RRT). We train the RRT model with official code provided by [14], and use all the same settings are as the provided one. The only difference is that we input the global descriptor extracted from CVNet-Global and local features extracted from the added local branch, instead of the features extracted by the pre-trained DELG model.

S5. Additional Qualitative Results

Additional qualitative results on ROxford5k-Hard+1M, RParis6k-Hard+1M, and the GLDv2-retrieval-test are shown in Fig. S6, Fig. S7, and Fig. S8, respectively. The results show that the proposed re-ranking method performs re-ranking robustly, even if the global descriptor matching results in misjudgment in situations involving challenges such as viewpoint change, occlusion, and truncation.

References

- [1] 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. 4, 5, 6, 7
- [2] 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. 5
- [3] 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. 5
- [4] 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. 6
- [5] 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. 5
- [6] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 5
- [7] Dmytro Mishkin, Jiri Matas, and Michal Perdoch. Mods: Fast and robust method for two-view matching. *Computer Vision and Image Understanding*, 2015. 7
- [8] 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. 5
- [9] 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, 4
- [10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 6
- [11] 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. 5
- [12] 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. 5
- [13] 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. 3
- [14] Fuwen Tan, Jiangbo Yuan, and Vicente Ordonez. Instance-level image retrieval using reranking transformers. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 6, 7
- [15] 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. 5
- [16] 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. 1
- [17] 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. 4
- [18] 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. 1

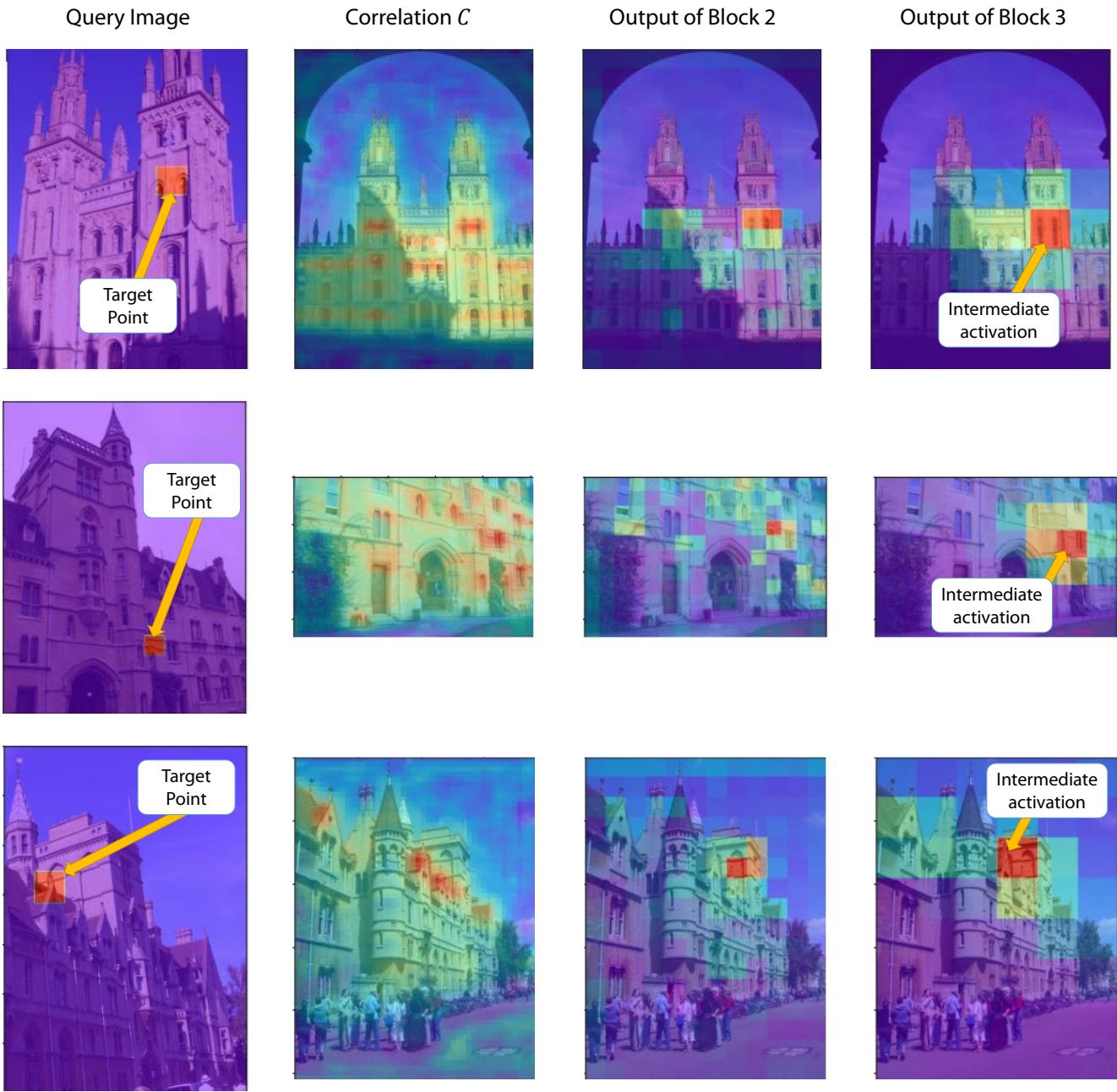


Figure S5. Additional Intermediate Feature Visualization.



Figure S6. Additional Qualitative Results on R(Oxford5k-Hard+1M) with R50-CVNet. The upper line is the global descriptor matching result and the lower line is 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. Our purpose is to visualize the difference between global descriptor matching and re-ranking, so we skip the results of the ranks that are correct in both the global descriptor matching and re-ranking processes.



Figure S7. Additional Qualitative Results on RParis6k-Hard+1M with R50-CVNet. The upper line is the global descriptor matching result and the lower line is 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. Our purpose is to visualize the difference between global descriptor matching and re-ranking, so we skip the results of the ranks that are correct in both the global descriptor matching and re-ranking processes.

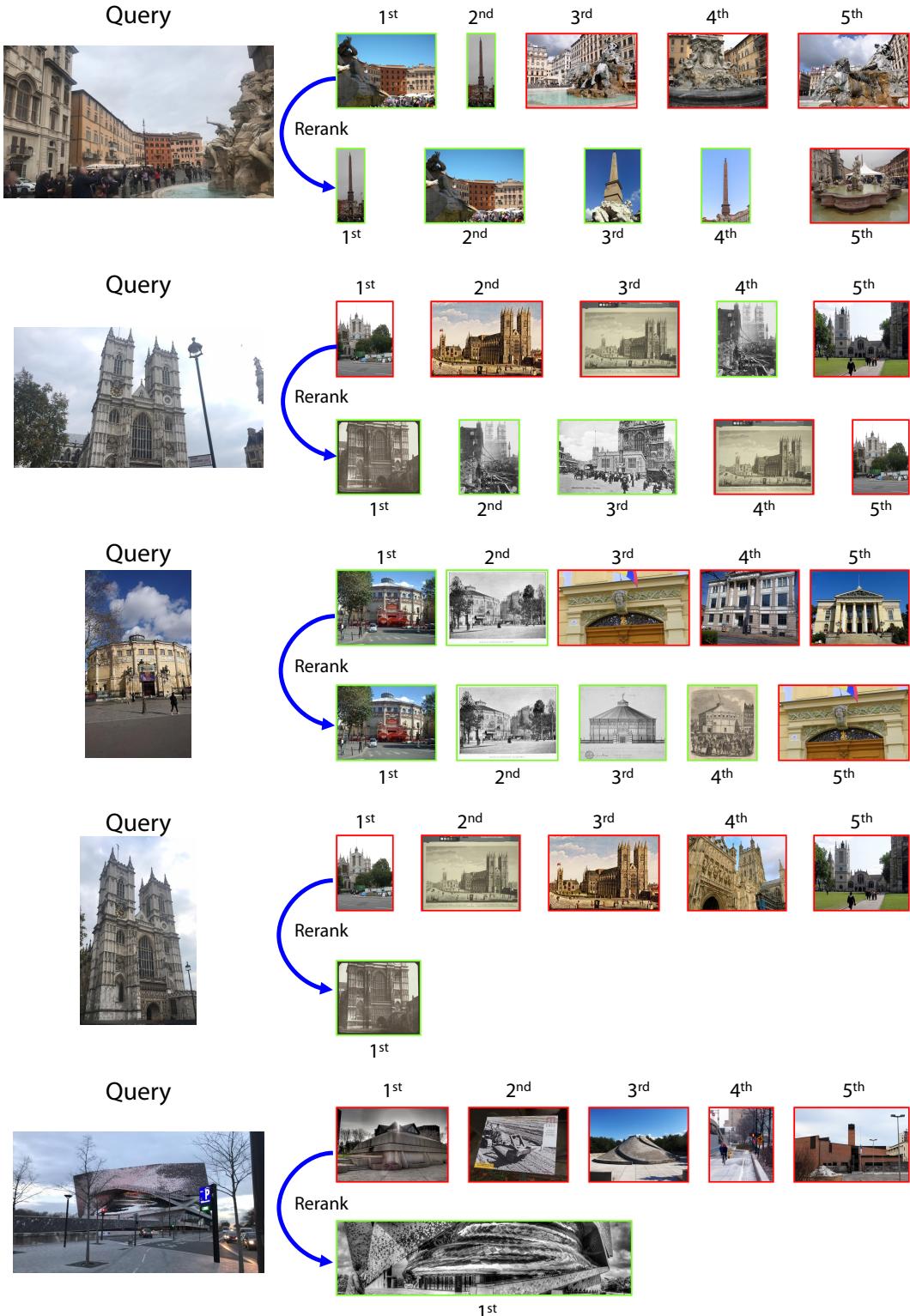


Figure S8. Qualitative Results on GLDv2-retrieval-test with R50-CVNet. The upper line is the global descriptor matching result and the lower line is the re-ranking result. Correct/incorrect results are marked with green/red borders, respectively. The last two queries each have only one positive sample, so we skip the results after the correct answer comes out.