Revisiting Self-Similarity: Structural Embedding for Image Retrieval

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

Abstract

Despite advances in global image representation, existing image retrieval approaches rarely consider geometric structure during the global retrieval stage. In this work, we revisit the conventional self-similarity descriptor from a convolutional perspective, to encode both the visual and structural cues of the image to global image representation. Our proposed network, named Structural Embedding Network (SENet), captures the internal structure of the images and gradually compresses them into dense self-similarity descriptors while learning diverse structures from various images. These self-similarity descriptors and original image features are fused and then pooled into global embedding, so that global embedding can represent both geometric and visual cues of the image. Along with this novel structural embedding, our proposed network sets new state-of-the-art performances on several image retrieval benchmarks, convincing its robustness to look-alike distractors. The code and models are available: https://github.com/sungonce/SENet.

1. Introduction

Content-based image retrieval is the task of searching for images with the same content present in the query image in the large-scale database. What across images represents the same content are two things: the visual properties and the geometrical structure, so comparing them well is the key to the image retrieval task. To achieve this goal, two image representation types have been extensively explored in many image retrieval solutions. The first one is local features [1–3, 5, 19–22, 24, 47] that comprise visual descriptors and spatial information about local regions of the image, and the other one is a global descriptor [3,11–13,18,24–26,28,33,43], also known as global embedding, that summarizes the local features of the entire image. In a general sense, the global descriptor loses spatial information of local features during the summarization process. Thus, many image retrieval solutions [3, 18, 24, 35, 36, 42] first retrieve coarse candidates with similar visual properties for the query using global embeddings (typically referred to as global retrieval) and further verify that coarse candidates have geometrically similar shapes to the query using local features (typically referred to as local feature re-ranking). This separation of tasks may sound reasonable at first glance. However, in fact, they miss the opportunities to perform robust retrieval by comparing structural information also in the global retrieval stage.

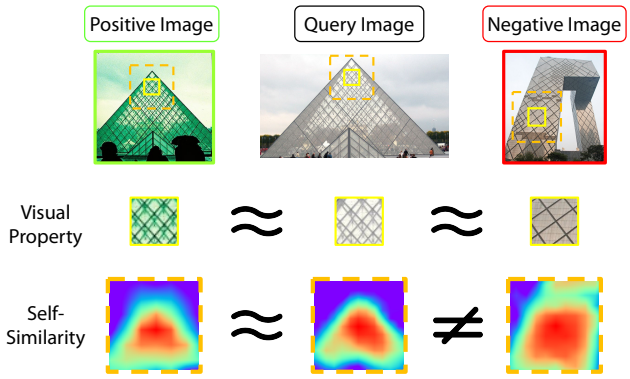

Figure 1. Images of the same content share both similar image properties and internal self-similarities. Our proposed networks leverage both visual features and self-similarity features and encode them to global embedding in an end-to-end manner.

In computer vision, a self-similarity descriptor [31] has long been used as a regional descriptor for matching images based on the aggregation of local internal structures. This work has shown its effectiveness in challenging matching problems even in situations where the visual properties of images are not shared at all (*e.g.* matching between drawing and photo domains). However, their self-similarity encoding process is neither learnable nor differentiable. And it also completely ignores visual properties, making it difficult to use directly for image retrieval tasks where visual properties are also valuable cues.

In this paper, we revisit the self-similarity descriptor in a convolutional manner and propose a novel global embedding network named ***S**tructural **E**mbedding **Net**work* (**SENet**). The proposed network captures the internal structures of the images and encodes them to self-similarity de-

*Corresponding author.

scriptors while learning diverse structures from various images. These self-similarity descriptors and original image features are fused and then pooled into global embedding, so that global embedding can represent both valuable geometric and visual cues of the image. All proposed modules of our networks are comprised of point-wise operations, enabling efficient descriptor encoding. Our proposed network sets state-of-the-art performance on several image retrieval benchmarks, convincing its robustness to look-alike distractors.

2. Related Work

Image retrieval. Image retrieval aims to search the database for images that contain the same content as in the query image. Representative solutions [3, 18, 24–26, 32, 35, 36, 45] in this task are mainly based on a coarse-to-fine approach: coarse one is the global retrieval with local feature aggregation that aggregates hand-crafted local features [11–13, 25, 26, 33] into a compact global embedding, and the fine one is re-ranking the coarse retrieval results with spatial verification [3, 24–26, 32, 36, 45] that verifies whether putative local feature correspondence constructs the rigid geometric relationship or not. In this local feature aggregation process, spatial information of local features is lost, and global embedding mainly reflects representative visual properties of images. With the advancement of deep learning, in recent studies [3, 23, 24, 28, 41, 44–46], local features have been replaced with intermediate features of CNNs, and aggregation methods have been replaced with various spatial pooling operations. DELF and DELG [24] presented a learning-based local and global feature using CNN, and [28] proposed a GeM pooling that aggregates local features through attentive pooling. SOLAR [23] showed a method of aggregation of attentive features based on self-attention inside image features, and DOLG [46] proposed a method of fusion of local and global features of an image. These changes boost image retrieval performances significantly; however, the geometric structure still has been mainly considered within the re-ranking stage, while the global embedding hardly reflects the structure of the image. Such discrimination makes global embeddings easy to fooled by look-alike distractors with similar visual patterns even if their shapes are quite different, ultimately bringing down the overall retrieval performance. To address this drawback, we propose a novel global embedding network to help representative visual properties and internal structures reflect well simultaneously in global embedding.

Self-similarity. Self-similarity is a structural representation of an image, indicating how similar a specific part of an image is to the entire image or its neighborhood region. For many previous studies, this self-similarity was mainly used for two purposes. The first is a self-similarity descriptor, which uses self-similarity as a robust local descriptor, and the second is self-attention, which enhances the attentive region with self-similarity as weight. In this work, we focus entirely on the self-similarity descriptor scheme to prove that the structural characteristics of images are helpful for image matching.

Self-similarity descriptor. The self-similarity descriptor [31], which uses self-similarity itself as a descriptor, is a classical image descriptor developed to robustly match images of the same content under different photometric conditions such as lighting changes, color variance, texture differences, or even domain variances. With the help of this structural consistency, self-similarity descriptor has been used in various fields such as image matching [4, 30, 31], visual correspondence [15, 16], few-shot classification [14], and video action recognition [17]. With the advance of deep learning, recent approaches [14, 15, 17] extract self-similarity descriptors from the intermediate feature map of CNN to help the network learn robust local and global representation. In this paper, inspired by the previous works, we revisit this self-similarity descriptor in terms of convolution and propose a self-similarity encoder that embeds structural properties in global embeddings while learning diverse structures within numerous images. Unlike previous studies [4, 15, 16, 30, 31] that used only self-similarity descriptors due to large differences in photometric conditions, both visual properties and geometric structures are valuable clues in image retrieval. Carefully considering this concern, we propose a feature fusion module capable of harmoniously fusing visual and structural features. Thanks to these powerful proposal modules, our proposed network is not easily fooled by distractors with similar appearances, such as color or texture, and more accurately finds exact matches that match both visual properties and geometric structures.

3. Structural Embedding Networks (SENet)

In this section, we revisit the conventional local self-similarity descriptor for image retrieval and introduce our proposed global embedding network, named **S**tructural **E**mbedding **Net**works (**SENet**).

3.1. Problem Setup and Overview

Global embedding network for image retrieval task aims to reflect representative information of an image to global embedding to match images in an efficient way. In many recent studies, global embedding networks achieve the stated purpose by extracting the intermediate feature map of Convolutional Neural Networks (CNNs) and pooling it into global embedding. This process is generally formulated as follows. Given an input image $\mathbf{I} \in \mathbb{R}^{3\times H_I \times W_I}$, the intermediate feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ is extracted from the intermediate layer of the backbone CNN network f. The extracted feature map $\mathbf{F}$ is aggregated into global embedding $\mathbf{z}$ through a global pooling operation. These global pooling operations naturally discard spatial information of

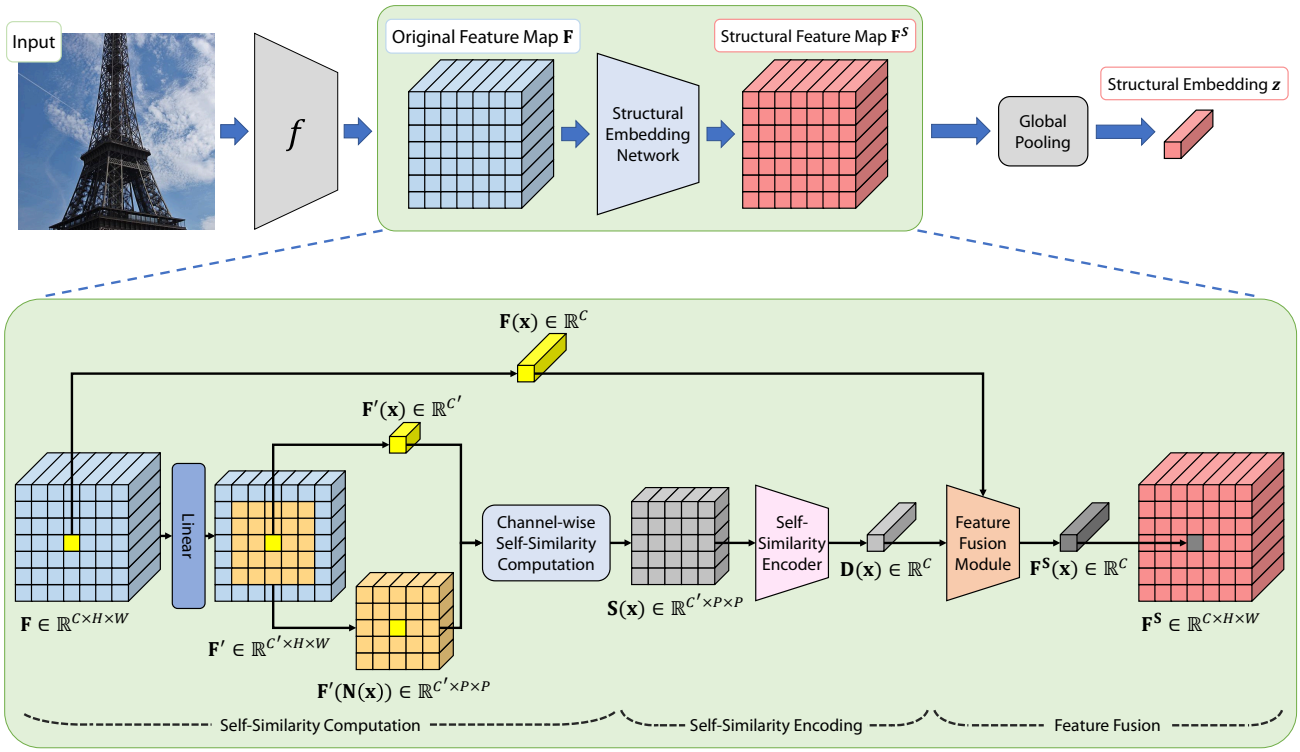

Figure 2. The overall structure of our proposed **S**tructural **E**mbedding **Net**work (**SENet**). Our proposed network captures the local structure patterns inside the image feature map $\mathbf{F}$ and fuses it with the original feature map, to extract the global embedding that contains both visual and structural cues of the input image. To this end, we organize our network into three parts: Self-Similarity Computation (SSC), Self-Similarity Encoder (SSE), and feature fusion Module (FFM). Since all proposed modules consist of pixel-wise operation. Therefore, for convenience, we show how single pixel position feature $\mathbf{F}(\mathbf{x})$ is encoded as a structural feature $\mathbf{F}^{\mathbf{s}}(\mathbf{x})$ in this figure.

the input feature map, so global embedding $\mathbf{z}$ hardly reflects the structural information about the input image, which is also an invaluable cue for image matching.

This paper focuses on making the global embedding better reflect structural information while preserving the original visual information. To this end, we introduce three modules comprised of Self-Similarity Computation (SSC, Sec. 3.2), Self-Similarity Encoder (SSE, Sec. 3.3), and Feature Fusion Module (FFM, Sec. 3.4). Our network computes self-similarity from the feature map with SSC, encodes it with pixel-wise self-similarity descriptors with SSE, and fuses them with the original feature map with FFM. With these modules, both the visual and structural information are well reflected in the feature map so that both visual and structural information is pooled well into global embedding. Owing to these structural considerations, SENet achieves superior retrieval ability. The overall architecture of SENet is outlined in Fig. 2.

3.2. Self-Similarity Computation (SSC)

First, we introduce the Self-Similarity Computation (SSC) module, which computes pixel-wise self-similarity $\mathbf{S}$ to extract this structural information inside the feature map $\mathbf{F}$. The SSC module takes the intermediate feature map $\mathbf{F}$ and transforms it to its projection $\mathbf{F}' \in \mathbb{R}^{C' \times H \times W}$ through the single linear layer to reduce computation complexity due to larger channel size while adding non-linearity to the original feature. Inside this projected feature map $\mathbf{F}'$, we compute channel-wise non-negative self-similarity $\mathbf{S} \in \mathbb{R}^{C' \times H \times W \times P \times P}$ for the every pixel position $\mathbf{x}$ and its surrounding region of size $P \times P$ using cosine similarity:

$$\mathbf{S}\left(\mathbf{c}, \mathbf{x}, \mathbf{d}\right) = \max\left(0, \frac{\mathbf{F}'(\mathbf{c}, \mathbf{x}) \cdot \mathbf{F}'(\mathbf{c}, \mathbf{x}+\mathbf{d})}{\|\mathbf{F}'(\mathbf{c}, \mathbf{x})\| \, \|\mathbf{F}'(\mathbf{c}, \mathbf{x}+\mathbf{d})\|}\right), \quad (1)$$

where $\mathbf{c} \in [1, C']$ is a index of channel dimension and $\mathbf{d} \in [-d_P, d_P] \times [-d_P, d_P]$ is relative position in the surrounding region of each pixel $\mathbf{x}$ with size of $P \times P$ such that $d_p = (P-1)/2$.

3.3. Self-Similarity Encoder (SSE)

In this subsection, we introduce a Self-Similarity Encoder (SSE), which encodes high-dimensional self-similarity into compact self-similarity descriptors while learning and analyzing diverse geometric structures from various images. SSE takes the dense channel-wise local self-similarity $\mathbf{S} \in \mathbb{R}^{C' \times H \times W \times P \times P}$ and gradually encodes it into dense self-similarity descriptor $\mathbf{D} \in \mathbb{R}^{C \times H \times W}$,

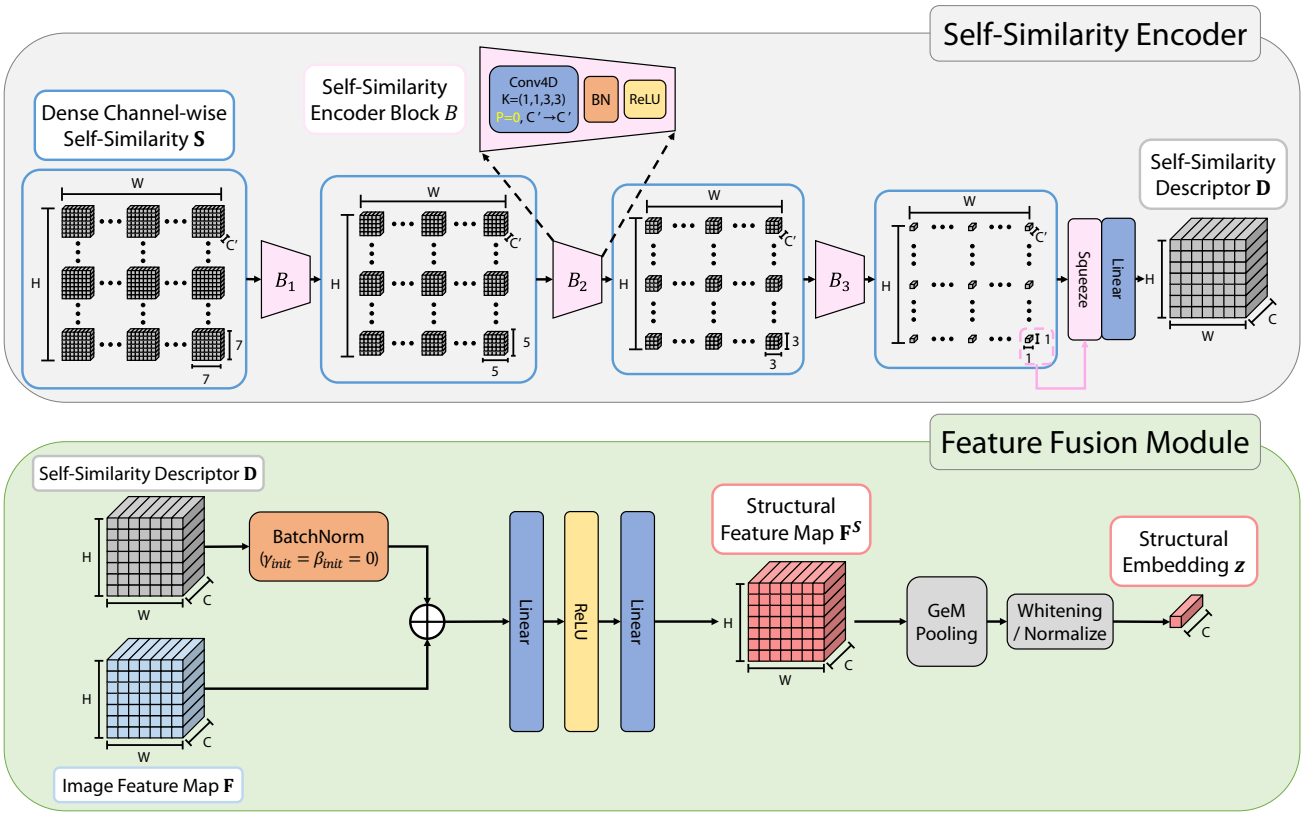

Figure 3. The detailed structure of the Self-Similarity Encoder (SSE) and Feature Fusion Module (FFM). Our proposed SSE encodes high-dimensional self-similarity into compact self-similarity descriptors, and FFM fuses self-similarity descriptors and the original feature map, to make the global embedding reflect both visual and structural information without breaking the original behavior of the base network.

which has the same spatial and channel size as original feature map $\mathbf{F}$. SSE is constructed with a convolution block sequence comprising 3×3 convolution along the self-similarity dimension side, batch normalization layer, and ReLU function. All operations in the encoder are pixel-wise (on the original image side), reducing the overall computational load. Following [14], the self-similarity encoder aggregates the self-similarity side by setting the padding value of the self-similarity side to zero, which reduces its spatial dimension from $P \times P$ to 1×1 while converting the raw self-similarity $\mathbf{S}$ to the self-similarity descriptor $\mathbf{D}$. Finally, a linear layer is followed to recover the channel size of the encoded self-similarity descriptor to the channel size of the original feature map $\mathbf{F}$. The detailed structure of our encoder is illustrated in Fig. 3.

3.4. Feature Fusion Module (FFM)

The self-similarity descriptor is a regional descriptor that reveals local geometric structures in the interest region while suppressing appearance variation inside it [31]. This self-similarity descriptor is helpful for image retrieval; however, it is not enough to use alone. Since the content inside the image has both visual and structural properties, we need to reflect both properties well in the global embedding step to improve search performance. Here, we introduce a Feature Fusion Module (FFM) that harmoniously fuses the self-similarity descriptors $\mathbf{D}$, which reflect the structural properties of the image and the original feature map $\mathbf{F}$, which reflect the visual properties of the image. First, we sum up the self-similarity descriptor and image feature map. Here, we add a batch normalization layer with the affine parameters (scale γ and bias β) initialized to zero before adding the self-similarity descriptor, following [7,40]. This initializing helps the self-similarity descriptor can be well fused to the original feature map without breaking the original behavior of the base network. Then the summed feature map feeds to a simple feed-forward layer [39], which comprises two linear layers and a ReLU function between them:

$$\mathbf{F^s}(\mathbf{x}) = \max\left(0, \left(\mathbf{F}(\mathbf{x}) + \mathbf{D}(\mathbf{x})\right) W_1 + b_1\right) W_2 + b_2, \quad (2)$$

where $\mathbf{F^s}$ is fused structural feature map, $\mathbf{x}$ is spatial pixel position, W_i and b_i are the weight and bias of i_{th} linear layer, respectively. In this module, all pixel positions are processed separately to reduce the overall computational load, also like SSE. This fused structural feature map $\mathbf{F^s}$ is finally aggregated into structural embeddings $\mathbf{z}$ via GeM pooling [28], whitening layer and L2 normalization. The

detailed structure of our fusion module is also illustrated in Fig. 3.

3.5. Training Objective

We use two loss functions that are widely used in the image retrieval field: classification loss [3, 18, 34, 43, 46] and contrastive loss [18, 23, 28].

Classification loss. For classification loss, following [18, 46], we adopt cosine classifier with CurricularFace [10] margin. Our classification loss $\mathcal{L}_{cls}$ is defined as:

$$\mathcal{L}_{cls} = -\log \frac{\exp(\mathcal{M}(W_{y(\mathbf{z})}^{\top}\mathbf{z}, 1)/\tau)}{\sum_{c=1}^{N}\exp(\mathcal{M}(W_c^{\top}\mathbf{z}, \mathbb{1}_{y(\mathbf{z})}^{i})/\tau)}, \quad (3)$$

where $y(\mathbf{z})$ is the ground-truth label of $\mathbf{z}$, W_c is the c_{th} class weight for the cosine classifier, τ is the temperature parameter, and the $\mathbb{1}_{y(\mathbf{z})}^{i}$ is the one-hot indicator whether the label index i and $y(\mathbf{z})$ is same or not. $\mathcal{M}$ is the function that adds the curricular margin to input logit s

$$\mathcal{M}(s, \mathbb{1}) = \begin{cases} \cos(\arccos(s) + m), & \text{if } \mathbb{1} = 1 \\ s, & \text{if } \mathbb{1} = 0, s < t \\ s(t + s), & \text{if } \mathbb{1} = 0, s > t \end{cases}, \quad (4)$$

where m is the margin value, and t is the moving average of query-positive logit.

Contrastive loss. For contrastive loss, following [18], we adopt MoCo [8]-style contrastive loss with CurricularFace [10] margin. Our classification loss $\mathcal{L}_{con}$ is defined as:

$$\mathcal{L}_{con} = -\mathop{\mathbb{E}}_{p \in P(\mathbf{z})} \log \frac{\exp\left(\mathcal{M}\left({d_p}^{\top}\mathbf{z}, 1\right)/\tau\right)}{\sum\limits_{i \in \{p, N(\mathbf{z})\}} \exp\left(\mathcal{M}\left({d_i}^{\top}\mathbf{z}, \mathbb{1}_{y(\mathbf{z})}^{y(d_i)}\right)/\tau\right)}, \quad (5)$$

where d_i is the i_{th} embedding inside the queue, and P and N are the index sets of samples in positive or negative relationships with embedding z in the queue, respectively. All other notations and parameters are the same as classification loss $\mathcal{L}_{cls}$, but are updated separately from those of $\mathcal{L}_{cls}$.

Total loss. The final loss $\mathcal{L}_T$ of our proposed network is either using only the classification loss $\mathcal{L}_{cls}$, which means $\mathcal{L}_T = \mathcal{L}_{cls}$, or using both the classification loss $\mathcal{L}_{cls}$ and contrastive loss $\mathcal{L}_{con}$, which means $\mathcal{L}_T = \alpha \cdot \mathcal{L}_{cls} + (1 - \alpha) \cdot \mathcal{L}_{con}$, where α is weight to fuse.

4. Experiments

4.1. Implementation Details

Training dataset. We use the Google Landmarks dataset v2-clean subset (referred to as GLDv2-clean) [42] to our training. GLDv2-clean consists of 1580470 images from 81313 landmarks with various landmarks.

Model design. We use ResNet-50 (R50) and ResNet-101 (R101) [9] as backbone networks and extract the intermediate feature map before the pooling operation (generally called '*conv5*' feature map), so the original channel size C is 2048 and we set the compressed channel size C' to 256. We set the self-similarity region size P is set to 7, so SSM has three encoder blocks as illustrated in Fig. 3.

Training details. We use random crop, aspect ratio distortion, and PCA color jittering augmentation to augment the training images. After augmentation, all training images are resized to 512×512 resolution. All proposed models are trained for 25 epochs. We use an SGD optimizer and set the initial learning rate of 5e-2, a batch size of 128, a momentum of 9e-1, and a weight decay of 1e-4. We adjust the learning rate for other batch sizes using linear scaling rules [7] to achieve a similar result. For the learning rate adjustment, we use a cosine learning rate schedule while warming up the learning process by setting the learning rate to 1/10 of the initial learning rate during the first epoch. For the CurricularFace margin of both classification loss $\mathcal{L}_{cls}$ and contrastive loss $\mathcal{L}_{con}$, we set the margin m to 0.15 and the temperature τ to $1/30$. For contrastive loss $\mathcal{L}_{con}$, we use a momentum network with a momentum of 0.999, a queue size of 73728, and fusion weight α of 0.5. We set the power of GeM Pooling as 3.0, which is fixed throughout the overall process. Finally, we present four models trained with two loss functions ($\mathcal{L}_{cls}$, $\mathcal{L}_{cls} + \mathcal{L}_{con}$) each for two backbones (R50, R101), respectively, to compare our proposed network with various previous solutions.

Embedding extraction. To match images in a multi-scale manner, we extract global embedding of three scales as follows: [0.7071, 1.0, 1.4142]. We get the final global embedding by L2-normalizing the average of the three L2-normalized embeddings. This three-scale global embedding extraction is used in most image retrieval studies [3, 6, 18, 28, 34, 35, 43] conventionally.

4.2. Evaluation Benchmarks

To verify the performance of our proposed network, we conduct experiments in Revisited Oxford ($\mathcal{R}$Oxf) [25, 27] and Revisited Paris ($\mathcal{R}$Par) [26, 27], which are representative benchmarks widely used in image retrieval studies. Both benchmarks have 70 query images and contain 4933 and 6322 database images, respectively. Additionally, we measure large-scale search performance by adding 1M distractor images (+1M) given by [27] to the database of both benchmarks. The retrieval performance on these two benchmarks is measured using a mean Average Precision (mAP).

4.3. Experimental Results

Comparison with the state-of-the-art models (Tab. 1). Tab. 1 shows the retrieval performance of our proposed network and the previous state-of-the-art image retrieval models tested on the $\mathcal{R}$Oxford and the $\mathcal{R}$Paris benchmarks, also

	Loss		Medium				Hard			
	$\mathcal{L}_{cls}$	$\mathcal{L}_{con}$	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M
(a) Local feature aggregation										
R101-HOW-VLAD (GLDv2-clean) [13, 38, 43]	✓		73.5	60.4	82.3	62.6	51.9	33.2	67.0	41.8
R101-HOW-ASMK (GLDv2-clean) [37, 38, 43]	✓		80.4	70.2	85.4	68.8	62.5	45.4	70.8	45.4
R50-FIRe-ASMK (SfM-120k) [41]		✓	**81.8**	66.5	85.3	67.6	61.2	40.1	70.0	42.9
R50-MDA-ASMK (SfM-120k) [44]		✓	**81.8**	**68.7**	83.3	64.7	**62.2**	**45.3**	66.2	38.9
R50-Token (GLDv2-clean) [43]	✓		80.5	68.3	**87.6**	**73.9**	62.1	43.4	**73.8**	**53.3**
R101-Token (GLDv2-clean) [43]	✓		**82.3**	**70.5**	**89.3**	**76.7**	**66.6**	**47.3**	**78.6**	**55.9**
(b) Global single-pass										
R101-GeM (SfM-120k) [29]		✓	65.3	46.1	77.3	52.6	39.6	22.2	56.6	24.8
R101-GeM-AP* (GLDv1) [28, 32]		✓	66.3	-	80.2	-	42.5	-	60.8	-
R101-GeM-ArcFace (GLDv2-clean) [42]	✓		74.2	-	84.9	-	51.6	-	70.3	-
R101-SOLAR (GLDv1) [23]		✓	69.9	53.5	81.6	59.2	47.9	29.9	65.5	33.4
R50-DELG (GLDv2-clean) [3]	✓		73.6	60.6	85.7	68.6	51.0	32.7	71.5	44.4
R101-DELG (GLDv2-clean) [3]	✓		76.3	63.7	86.6	70.6	55.6	37.5	72.4	46.9
R50-DOLG‡ (GLDv2-clean) [46]	✓		78.6	68.9	87.5	76.7	58.2	44.1	73.7	56.2
R101-DOLG‡ (GLDv2-clean) [46]	✓		79.5	72.1	89.7	80.3	59.5	47.8	78.1	61.5
R101-GLAM (GLDv2-clean) [34]	✓		78.6	68.0	88.5	73.5	60.2	43.5	76.8	53.1
R50-CVNet-Global (GLDv2-clean) [18]	✓	✓	**81.0**	**72.6**	**88.8**	**79.0**	**62.1**	**50.2**	**76.5**	**60.2**
R101-CVNet-Global (GLDv2-clean) [18]	✓	✓	**80.2**	**74.0**	**90.3**	**80.6**	**63.1**	**53.7**	**79.1**	**62.2**
(c) Ours										
R50-SENet (GLDv2-clean)	✓		81.4	72.9	**90.5**	79.0	62.3	48.7	**80.3**	**59.9**
R50-SENet (GLDv2-clean)	✓	✓	**81.9**	**74.2**	90.0	**79.1**	**63.0**	**52.0**	78.1	**59.9**
R101-SENet (GLDv2-clean)	✓		80.0	72.5	91.6	82.1	61.7	49.2	82.2	64.6
R101-SENet (GLDv2-clean)	✓	✓	**82.8**	**76.1**	**91.7**	**83.6**	**66.0**	**55.7**	**82.8**	**67.8**

Table 1. **Comparison with the state-of-the-art models.** evaluated performances on Revisited Oxford ($\mathcal{R}$Oxf) and Revisited Paris ($\mathcal{R}$Par) with adding 1M distractors experiments (+1M). The best scores for each group and backbone are presented as **boldfaced** text. $\star$ denotes using AP-Loss. ‡ denotes reproduced model with the official model and weights.

with their 1 million distractors. In this table, we divided the previous global embedding models into two groups: (a) Local feature aggregation and (b) Global single-pass. *(a) Local feature aggregation.* Our proposed network shows better overall performance than solutions using classical aggregators such as VLAD [12] or ASMK [37]. However, they consume more extract and matching time due to the aggregator. Recently, a learnable aggregator named Token [43] has been proposed and sets a new state-of-the-art performance with reasonable extraction latency and matching time. However, due to the limitation that structural characteristics cannot be considered, it shows a substantial performance decrease when a 1M distractor is added. Our proposed method leads the performance by up to 8.7% ($\mathcal{R}$Par-Hard+1M) when adding 1M distractor experiments in the setting using the same loss function (R101-SENet-$\mathcal{L}_{cls}$). *(b) Global single-pass.* From some point of view, our proposed networks belong to this group. In this group, our proposed networks outperform existing solutions in most experiments regardless of the setting. Among the solutions using classification loss only, the state-of-the-art solution is DOLG[1] [46]. DOLG is a method that has learned four times the average learning time of other solutions, but the proposed network outperforms it on all measures. The state-of-the-art solution that uses both classification loss and contrast loss is CVNet-Global [18], which is a global backbone network of CVNet. In the same loss function setting, our network also outperforms CVNet-Global with a large gap in all measures by up to 5.6% on $\mathcal{R}$Par-Hard+1M with R101-SENet-$\mathcal{L}_{cls}$ & $\mathcal{L}_{con}$ model.

Comparison with reproduced models (Tab. 2). Many image retrieval models are trained and tested based on their environment settings (*e.g.* training dataset, loss function, multi-scale extraction, pooling method). For a fair comparison, we reproduced four representative image retrieval models (DELG [3], SOLAR [23], DOLG [46], and CVNet [18]) with the same settings as Sec. 4.1. Our proposed networks surpass other reproduced models in all measures for all models and settings. We also check the extraction time and memory costs for all reproduced models and ours. DELG and CVNet, which use almost pure ResNet structure, have the lowest latency and number of parameters, and SENet has 40% more parameters and 14% more latency than them. Since all the processes of the contributed modules of SENet are pixel-wise operations, extraction latency is in a reasonable time, which is slightly faster than SOLAR, despite the increased number of parameters.

[1] We reproduced the DOLG with the official model and weights due to the misreported performance of the original paper.

model	Medium				Hard				Params	Time
(R50)	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M	(M)	(ms)
(a) Use only classification loss $\mathcal{L}_{cls}$										
DELG†	78.6	70.7	89.5	77.4	58.8	44.8	77.9	57.7	**27.7**	**8.82**
SOLAR†	79.0	70.1	89.8	78.5	59.0	44.9	78.2	59.0	37.2	10.21
DOLG†	78.7	70.9	88.9	76.6	57.9	45.5	75.8	55.9	30.9	9.32
SENet	**81.4**	**72.9**	**90.5**	**79.0**	**62.3**	**48.7**	**80.3**	**59.9**	38.9	10.05
(b) Use both classification loss $\mathcal{L}_{cls}$ and contrastive loss $\mathcal{L}_{con}$										
CVNet†	79.7	72.4	89.5	78.9	60.5	49.6	77.6	59.5	**27.7**	**8.82**
SENet	**81.9**	**74.2**	**90.0**	**79.1**	**63.0**	**52.0**	**78.1**	**59.9**	38.9	10.05

Table 2. **Comparison with reproduced state-of-the-art models**. † denotes reproduced model with the same setting as SENet.

SSE	FFM		Medium				Hard			
	BN$_{init}$	FFL	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M
			78.6	70.7	89.5	77.4	58.8	44.8	77.9	57.7
✓			79.2	71.2	89.2	76.7	59.3	45.2	76.8	56.3
✓	✓		79.8	71.9	90.0	78.0	60.7	47.2	78.5	58.3
✓	✓	✓	**81.4**	**72.9**	**90.5**	**79.0**	**62.3**	**48.7**	**80.3**	**59.9**

Table 3. **Effect of the proposed modules.**

4.4. Ablation Studies

We conduct extensive ablation studies to analyze our proposed network. All ablation studies are performed with ResNet-50 backbone with classification loss only setting.

Effect of the proposed modules (Tab. 3). We conduct ablation studies on the proposed modules to demonstrate their efficacy. When using the self-similarity encoder (SSE) solely, the proposed method showed slightly better performance than the baseline but showed instability in the initial convergence. For smooth initial convergence, we added a batch normalization layer with scale and bias initialized to zero (BN$_{init}$) before summing two features. Hence, the module trained stably and showed slightly higher performance than before. Finally, when adding a simple pixel-wise feed-forward layer (FFL), two different characteristic features were fused harmoniously and showed the best performance.

Analysis on input feature layer (Tab. 4). We conduct ablation studies on which layers it is effective to extract structural information from. As shown in Tab. 4, extracting structural information from the *conv5* layer showed higher performance than the *conv4* layer. Even though *conv5*'s output feature has a lower resolution than *conv4*'s output feature, it enables the network to embed meaningful self-similarity by exploiting high-level semantic features. Therefore, we use the intermediate feature map from *conv5* layer to extract internal self-similarities.

Analysis on self-similarity region size (Tab. 5). We further analyze the effect of self-similarity region size P. In this experiment, SENet performs better than the baseline with all kernel size settings. Interestingly, the performance in the $\mathcal{R}$Paris benchmark is better in the high region size setting, and the performance in the $\mathcal{R}$Oxford benchmark is better in a relatively small region size setting. This is presumed to be due to a scale difference in benchmark images. Finally, we choose $P = 7$, which shows fine performance with moderate memory and latency burden.

model	Medium				Hard			
(R50, $\mathcal{L}_{cls}$)	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M
baseline	78.6	70.7	89.5	77.4	58.8	44.8	77.9	57.7
SENet (*conv4*)	79.2	71.1	89.9	77.5	59.5	46.0	78.2	57.2
SENet (*conv5*)	**81.4**	**72.9**	**90.5**	**79.0**	**62.3**	**48.7**	**80.3**	**59.9**

Table 4. **Ablation experiments on the input feature layer.**

model	Medium				Hard			
(R50, $\mathcal{L}_{cls}$)	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M
baseline	78.6	70.7	89.5	77.4	58.8	44.8	77.9	57.7
SENet ($P = 5$)	80.6	71.7	90.6	78.4	61.5	46.6	79.9	58.5
SENet ($P = 7$)	**81.4**	**72.9**	90.5	79.0	**62.3**	**48.7**	**80.3**	**59.9**
SENet ($P = 9$)	80.1	71.0	**90.7**	**79.1**	60.0	44.8	**80.3**	**59.9**

Table 5. **Ablation experiments on self-similarity region size P.**

Visualize the effect of proposed modules (Fig. 4). To verify that our proposed module is producing the intended effect properly, we conduct a visualization using the output of each module. We selected the positive image and the hard-negative image of the existing solutions for the query image, calculated the query-positive and query-negative feature similarity, and visualized it for a single spatial point $\mathbf{x}$ of the query image. Since the similar visual property of target point $\mathbf{x}$ exists in both positive image $\mathbf{I_p}$ and negative image $\mathbf{I_n}$, both query-positive similarity $\mathbf{S_c}(\mathbf{F_q}(\mathbf{x}), \mathbf{F_p})$ and similarity $\mathbf{S_c}(\mathbf{F_q}(\mathbf{x}), \mathbf{F_n})$ between their original feature map $\mathbf{F}$ show high values in some region. However, since the similar geometrical structure of the target point $\mathbf{x}$ exists in positive image $\mathbf{I_p}$ but does not exist in negative image $\mathbf{I_n}$, the query-positive similarity $\mathbf{S_c}(\mathbf{D_q}(\mathbf{x}), \mathbf{D_p})$ between self-similarity descriptors $\mathbf{D}$ shows high values in the corresponding region, while there are few high values in the query-negative similarity $\mathbf{S_c}(\mathbf{D_q}(\mathbf{x}), \mathbf{D_n})$. Finally, the original features $\mathbf{F}$ and self-similarity descriptor $\mathbf{D}$ are fused to structural feature $\mathbf{F^s}$, raising the similarities where both visual and structural cues form a consensus and diminishing the similarities that do not, as shown in $\mathbf{S_c}(\mathbf{F^s_q}(\mathbf{x}), \mathbf{F^s_p})$ and $\mathbf{S_c}(\mathbf{F^s_q}(\mathbf{x}), \mathbf{F^s_n})$.

5. Discussion

Qualitative results. Example qualitative results are shown in Fig. 5. Despite advanced feature representation, previous solutions that do not consider the internal structures of images are easily fooled by look-alike distractors. Our proposed network captures structural information and reflects them in global embeddings, thereby finding the correct answer more precisely considering the structure of the image even in the global search stage.

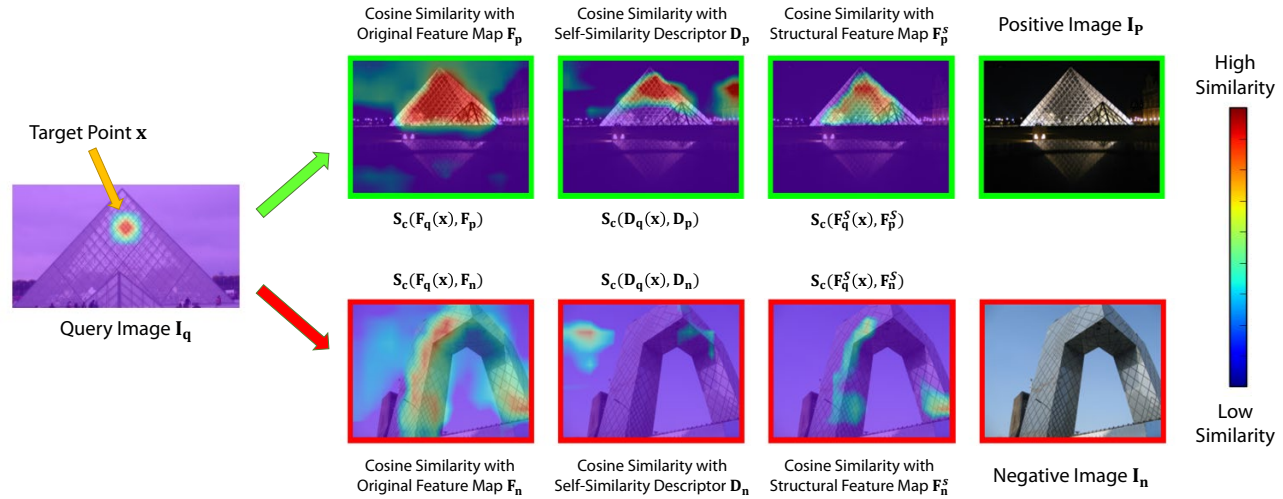

Figure 4. Example visualization of the intermediate feature similarity between query-positive and query-hard negative images. Our network enhances the similarity where the visual and structural cues form a consensus and diminishes other parts. $S_c(\cdot,\cdot)$ denotes cosine similarity between two inputs. All features are extracted using R50-SENet-$\mathcal{L}_{cls}$ model.

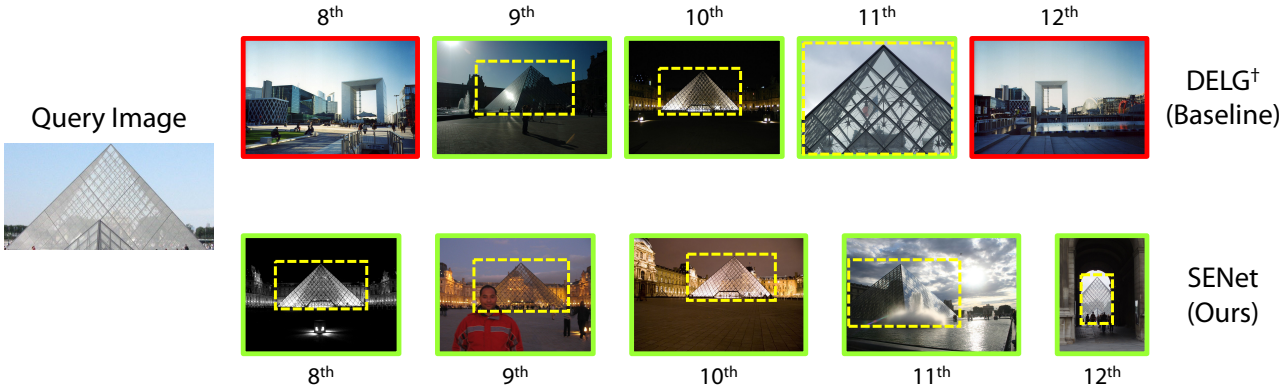

Figure 5. Example qualitative results with R50-DELG† and R50-SENet-$\mathcal{L}_{cls}$. The upper line results from R50-DELG†, and the lower line results from R50-SENet. Correct and incorrect answers are marked with green / red borders around the image, respectively. The yellow dotted line indicates the area of the positive image that overlaps the query.

Limitations and future work. Although our proposed network shows promising performance improvements, our proposed network still has weaknesses against the scale and the structural changes caused by image resolution and viewpoints. To solve this problem, we design an end-to-end learnable self-similarity encoder, which learns various structures from various images, but there is still room for performance improvement. Our future work aims to design a model that can enhance the consensus between the same content images with large scale/viewpoint differences.

6. Conclusion

In this paper, we present a novel framework that leverages the internal structures of images to reflect structural information well in global embeddings. To this end, we propose two modules. First, we propose the self-similarity encoding module, which analyzes the internal structures of images and encodes them to self-similarity descriptors in an end-to-end manner. We also propose the feature fusion module, to fuse visual and structural information harmoniously without breaking the original behavior of the base structure. The significant performance improvements on several representative benchmarks and intensive ablation studies demonstrate that the internal structures of images are also invaluable cues for image retrieval.

Acknowledgements. This work was supported in part by the KIST Institutional Program (Project No. 2E32271-23-078). This work was also supported in part by the Robot Industry Technology Development Project, 20023455, Development of Cooperate Mapping, Environment Recognition and Autonomous Driving Technology for Multi Mobile Robots Operating in Large-scale Indoor Workspace, funded by the Ministry of Trade, Industry & Energy (MOTIE, Republic of Korea).

References

[1] Axel Barroso-Laguna, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Key. net: Keypoint detection by handcrafted and learned cnn filters. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5836–5844, 2019. 1

[2] 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. 1

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

[4] Thomas Deselaers and Vittorio Ferrari. Global and efficient self-similarity for object classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1633–1640. IEEE, 2010. 2

[5] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[6] 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. 5

[7] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 4, 5

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

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[10] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5901–5910, 2020. 5

[11] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 304–317. Springer, 2008. 1, 2

[12] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3304–3311. IEEE, 2010. 1, 2, 6

[13] 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. 1, 2, 6

[14] Dahyun Kang, Heeseung Kwon, Juhong Min, and Minsu Cho. Relational embedding for few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8822–8833, October 2021. 2, 4

[15] Seungryong Kim, Dongbo Min, Bumsub Ham, Stephen Lin, and Kwanghoon Sohn. Fcss: Fully convolutional self-similarity for dense semantic correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(3):581–595, 2019. 2

[16] Seungryong Kim, Seungchul Ryu, Bumsub Ham, Junhyung Kim, and Kwanghoon Sohn. Local self-similarity frequency descriptor for multispectral feature matching. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 5746–5750, 2014. 2

[17] Heeseung Kwon, Manjin Kim, Suha Kwak, and Minsu Cho. Learning self-similarity in space and time as generalized motion for video action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13065–13075, October 2021. 2

[18] Seongwon Lee, Hongje Seong, Suhyeon Lee, and Euntai Kim. Correlation verification for image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5374–5384, June 2022. 1, 2, 5, 6

[19] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004. 1

[20] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2527–2536, 2019. 1

[21] 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. 1

[22] Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Repeatability is not enough: Learning affine regions via discriminability. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 284–300, 2018. 1

[23] Tony Ng, Vassileios Balntas, Yurun Tian, and Krystian Mikolajczyk. Solar: second-order loss and attention for image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 253–270. Springer, 2020. 2, 5, 6

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

[25] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2007. 1, 2, 5

[26] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008. 1, 2, 5

[27] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5706–5715, 2018. 5

[28] 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. 1, 2, 4, 5, 6

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

[30] Amin Sedaghat and Hamid Ebadi. Distinctive order based self-similarity descriptor for multi-sensor remote sensing image matching. *ISPRS Journal of Photogrammetry and Remote Sensing*, 108:62–71, 2015. 2

[31] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007. 1, 2, 4

[32] Oriane Siméoni, Yannis Avrithis, and Ondrej Chum. Local features and visual words emerge in activations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11651–11660, 2019. 2, 6

[33] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, volume 3, pages 1470–1470. IEEE Computer Society, 2003. 1, 2

[34] Chull Hwan Song, Hye Joo Han, and Yannis Avrithis. All the attention you need: Global- local, spatial-channel attention for image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2754–2763, 2022. 5, 6

[35] Fuwen Tan, Jiangbo Yuan, and Vicente Ordonez. Instance-level image retrieval using reranking transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 5

[36] Marvin Teichmann, Andre Araujo, Menglong Zhu, and Jack Sim. Detect-to-retrieve: Efficient regional aggregation for image search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5109–5118, 2019. 1, 2

[37] Giorgos Tolias, Yannis Avrithis, and Hervé Jégou. To aggregate or not to aggregate: Selective match kernels for image search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2013. 6

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

[39] 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. 4

[40] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 4

[41] Philippe Weinzaepfel, Thomas Lucas, Diane Larlus, and Yannis Kalantidis. Learning super-features for image retrieval. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 2, 6

[42] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2575–2584, 2020. 1, 5, 6

[43] Hui Wu, Min Wang, Wengang Zhou, Yang Hu, and Houqiang Li. Learning token-based representation for image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 2703–2711, Jun. 2022. 1, 5, 6

[44] Hui Wu, Min Wang, Wengang Zhou, and Houqiang Li. Learning deep local features with multiple dynamic attentions for large-scale image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11396–11405, 2021. 2, 6

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

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

[47] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 467–483. Springer, 2016. 1