ULTRON: Unifying Local Transformer and Convolution for Large-scale Image Retrieval

Minseong Kweon[1][0000−0001−8028−2930] and Jinsun Park[2,3⋆][0000−0002−2296−819X]

[1] School of Mechanical Engineering, Pusan National University, Republic of Korea
[2] School of Computer Science and Engineering, Pusan National University, Republic of Korea
[3] Center for Artificial Intelligence Research, Pusan National University
{wou1202,jspark}@pusan.ac.kr

Abstract. In large-scale image retrieval, the primary goal is to extract discriminative features and embed them into global image representations. Previous methods based on CNNs effectively learn local features and create robust representations, leading to strong performance. Transformers that excel in learning global context, however, often struggle to extract fine details and therefore do not perform well in large-scale landmark recognition. In this paper, we propose a novel hybrid architecture named ULTRON, which combines transformer blocks with local self-attention and a convolution-based encoder. Our local transformer block contains an advanced self-attention mechanism that enhances the spatial context awareness of key features and updates the value features by considering broader information within fixed-size regional windows. In addition, we have designed a channel-wise dilated convolution that adjusts dilation per channel, enabling effective multiscale feature learning while robustly capturing local features. We focus on learning local contexts throughout the entire network and effectively blending these contexts in the attention-based pooling process. This approach generates a powerful global representation that includes local information, relying solely on classification loss without requiring additional modules to capture local features. Experimental results demonstrate that our model outperforms previous works due to effectively embedding local features into a global representation.

Keywords: Image retrieval · Landmark recognition · Local self-attention.

1 Introduction

Content-based image retrieval refers to searching large databases to find images that match a query image. Specifically, instance-level image retrieval involves searching for visually similar images and identifying specific landmarks, buildings, or objects to find images containing matching entities. The key challenge in understanding the complex patterns of various locations and landmarks lies in

⋆ Corresponding author.

extracting distinctive and meaningful local features while achieving a compact global representation. In the early stages of feature extraction, various hand-crafted techniques [22, 5, 25, 44, 15, 39] were developed. Over the past decade, these traditional methods have evolved into techniques utilizing CNNs [33, 4] for object recognition. Notably, deep local features [24] extracted from dense feature maps generated by CNNs are essential for discerning fine-grained details and variations in images. These features are critical in image retrieval for geometric verification [10, 2] to determine whether two images are correctly matched. Recently, two-stage retrieval methods [6, 37, 19] have been proposed, where global retrieval is initially performed using deep global features, followed by refining a few errors based on local features. These approaches have achieved state-of-the-art performance in instance-level image retrieval. However, in real-time applications such as visual localization or SLAM, the two-stage method often increases system complexity and can introduce significant latency during inference. DOLG [48] employed geometric orthogonal fusion to address these issues, while other approaches [23, 34, 36] utilized attention mechanisms to integrate local and global features. Recent studies [18, 49] have been designed to embed richer detailed information into global features effectively.

While these efforts focus on improving instance-level image retrieval, other computer vision tasks have seen success with models based on the Vision Transformer (ViT) [8]. However, these advancements have not shared the same success in large-scale image retrieval tasks [9]. A few studies have introduced transformer-based [26] and hybrid [35] models that integrate atrous convolution [7] with ViT to recognize local features and embed them into global features. Despite achieving excellent performance, these approaches rely on large deep networks, resulting in slow inference times and high computational costs.

In this paper, we propose a **U**nifying **L**ocal **TR**ansformer and c**ON**volution network (**ULTRON**), a novel hybrid model that leverages both convolution and local self-attention blocks. We also introduce **S**patial **C**ontext-**A**ware **L**ocal **A**ttention (**SCALA**) as an advanced local self-attention technique to enhance spatial interdependency in our transformer blocks. Moreover, we design a **C**hannel-wise **D**ilated **Conv**olution (**CDConv**) that adjusts dilation rates across different channels to exploit multiscale information in convolution-based encoder blocks. The proposed method has achieved state-of-the-art performance on the revisited Oxford ($\mathcal{R}$Oxf) and Paris ($\mathcal{R}$Par) datasets with one million distractors ($\mathcal{R}$Oxf+1M and $\mathcal{R}$Par+1M) under the hard configuration [30].

2 Related Work

2.1 Image Retrieval

Image retrieval identifies matching images from a database given a query, using similarity measures. Early methods focused on hand-crafted local features [22, 5, 25, 44] and compact descriptors from feature aggregation [15, 39]. In addition, local feature matching with RANSAC [10] is used to re-rank the initial retrieval results, leveraging geometric verification for better alignment [27, 28].

With the advent of deep learning, features extracted from deep neural networks replaced hand-crafted ones [4, 11, 24], while methods like ASMK [40], R-ASMK [38], and Token [46] advanced local feature aggregation. Pooling methods such as SPoC [3], MAC [42], and GeM [31] generated more compact global descriptors, leading to deep global feature-based retrieval. Attention-based models like SOLAR [23] and GLAM [34] improved global representations by focusing on important regions. Alongside the progress in global representations, two-stage retrieval methods incorporating re-ranking as post-processing, such as DELG [6] and RRT [37], also evolved. However, due to the computational demands of such methods, global representation-based retrieval approaches have gained popularity for real-time applications, offering lower memory usage and improved efficiency [48, 36]. Recently, SENet [18], SpCa [49], and CFCD [50] pursued capturing richer details and enhancing feature discrimination to create more powerful global representations.

In contrast, our ULTRON is a single-stage model that effectively learns regional context throughout the entire network using only classification loss. This enables instance image retrieval through a single global pass using an end-to-end approach without additional local feature enhancement modules or pairwise loss functions, thereby reducing the complexity during the training process, while still retaining robust global representations that preserve local features.

2.2 Vision Transformer for Image Retrieval

Transformers have proven effective in re-ranking tasks [37], outperforming traditional methods like geometric verification and query expansion [12]. However, in terms of feature extraction, CNN backbone networks [24, 6, 48] have consistently achieved state-of-the-art results in large-scale landmark retrieval, whereas vanilla ViTs [9] have struggled to deliver satisfactory performance. In recent times, ViT-based [26] and hybrid [35] models have been proposed, showing improved results. Nevertheless, these networks integrate an additional module similar to ASPP [7] to learn local features, making them not significantly different from existing CNN-based architectures such as DOLG [48]. Moreover, they are resource-intensive due to the large backbone networks, such as ViT-B [8].

To address this, we have designed a compact hybrid model with fewer parameters than a fully ViT-based backbone, while still maintaining high performance. The proposed hybrid model, combining CNNs and ViTs, projects local features into a global representation, setting itself apart from previous approaches that relied on additional modules and post-processing to capture local context.

2.3 Local Self-Attention Mechanism

The self-attention mechanism in ViT [8] effectively captures long-range dependencies and learns global context but lacks awareness of local features. Additionally, it exhibits linear time complexity concerning the embedding dimension and quadratic complexity relative to the input size, leading to significant computational demands. To mitigate these issues, SASA [32] applies self-attention

within predefined window sizes rather than across the entire global feature map. Subsequently, Swin Transformer [21] introduced a shifted-window self-attention mechanism that alternates between two partition configurations in successive transformer blocks. This approach efficiently computes non-overlapping features while establishing connections between them. The more advanced form of local self-attention, namely Neighborhood Attention [13], focuses exclusively on self-attention within nearby features, along with efficient window-sized matrix multiplication methods. These local self-attention mechanisms enhance local feature recognition compared to traditional self-attention methods but often compromise the ability to capture long-range dependencies.

To improve the inter-dependency between features in the existing local self-attention mechanism, our model facilitates an extension of spatial context awareness during key-query interaction by considering a broader receptive field from the query's perspective, without expanding the feature updating area. As a result, the model can identify the importance of nearby features across a wider area for the given query.

3 Method

Fig. 1 illustrates the overall architecture of the proposed ULTRON. Let H, W, and C denote the height, width, and number of channels, respectively. Given a query image $\mathbf{I} \in \mathbb{R}^{C \times H \times W}$, ULTRON generates a deep global feature $\mathbf{d} \in \mathbb{R}^{D}$, which also integrates fine-grained details. Here, D indicates the final embedding dimension. In the shallow layers, we utilize Channel-wise Dilated Convolution, a refined convolution block that adapts the size of the receptive fields according to the importance of each channel. In the deep layers, a novel local self-attention mechanism called Spatial Context-Aware Local Attention is employed. This mechanism calculates the attention score between broadened key and fixed query features, ensuring strong local context awareness and mid-range interdependency. Finally, instead of a class token, we apply an attention-based global pooling technique [29] to represent a single global vector from the deep spatial feature map. The following sections provide a detailed explanation of each process.

3.1 Channel-wise Dilated Convolution

In the initial two stages, we design Channel-wise Dilated Convolution to dynamically adjust the receptive field of the spatial feature map based on channel significance, functioning as a multiscale encoder. Specifically, channel attention scores are computed, enabling smaller kernels with larger dilations to be applied to less important channels, while larger kernels with smaller dilations are assigned to channels with higher scores. We calculate the channel attention a_c according to ECA [43], which is defined as follows:

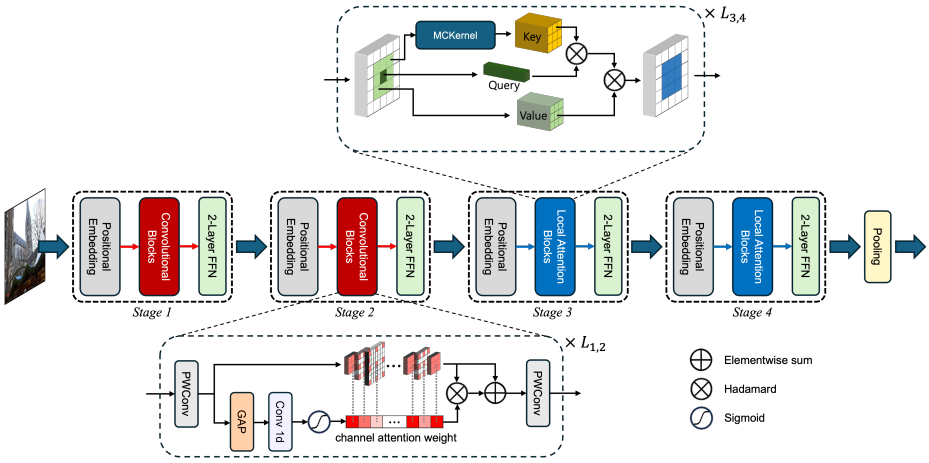


Fig. 1: The overall architecture of the proposed ULTRON. The first and second stages utilize the proposed CDConv, while the third and fourth stages employ the proposed SCALA.

$$a_c = \texttt{sigmoid}\left(\sum_{k=1}^{5} w_k \cdot \left(\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{c+k-P,i,j}\right)\right) \tag{1}$$

where c, P, k, w_k, a_c, i, and j represent the channel, padding, kernel index, weight of the 1D convolution, attention score of c-th channel, and spatial positions, respectively. Based on the calculated channel attention, dilation values for channel-wise convolutional layers are assigned.

Let τ_1 and τ_2 be the thresholds for the attention scores, and δ_1 and δ_2 denote the dilation values. Then, the dilation value d_c for the c-th channel, based on the attention weight a_c, can be defined as:

$$d_c = \begin{cases} 1 & \tau_1 < a_c, \\ \delta_1 & \tau_2 < a_c \leq \tau_1, \\ \delta_2 & a_c \leq \tau_2. \end{cases}$$

The CDConv is computed according to the following equation:

$$\text{CDConv}_{c,i,j}(X) = \sum_{m=-K}^{K}\sum_{n=-K}^{K} w_{c,m,n} \cdot X_{c,i+m\cdot d_c,j+n\cdot d_c}, \tag{2}$$

where m, n, and $w_{c,m,n}$ represent the row and column indices in the kernel, and the weight of kernel for channel c, respectively. The term d_c denotes the dilation rate determined by the attention score a_c. The feature map computed by CDConv is added to the feature map multiplied by the channel attention scores. Before and after performing this process, pointwise convolution with a

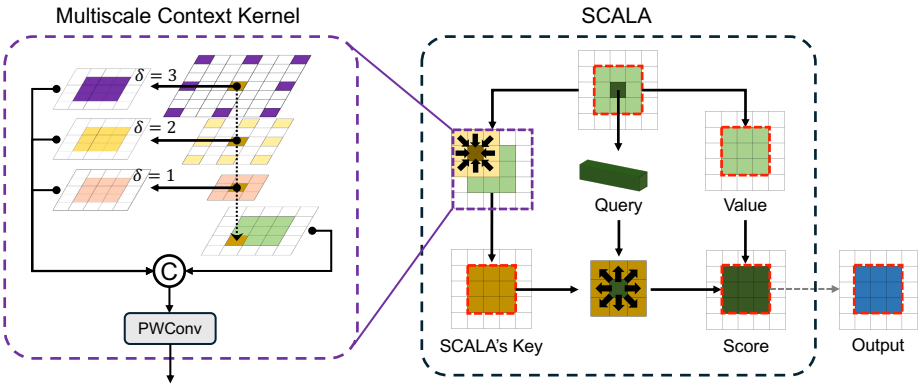


Fig. 2: Overview of the Spatial Context-Aware Local Attention. The red dashed rectangle represents the fixed window area where local self-attention is performed, while the purple dashed rectangle indicates the multiscale context kernel. The gray dashed lines denote the linear projection, and the black solid lines represent the progression of SCALA.

kernel size of 1×1 is applied to enhance computational efficiency. This structure can be interpreted as an advanced multiscale approach, enhancing the flexibility and efficiency of traditional convolution operations. Unlike the simple pointwise-depthwise-pointwise convolutions used in MobileNet [14], our method leverages adaptive dilation rates informed by attention mechanisms, allowing for dynamic receptive field adjustments tailored to the importance of different channels.

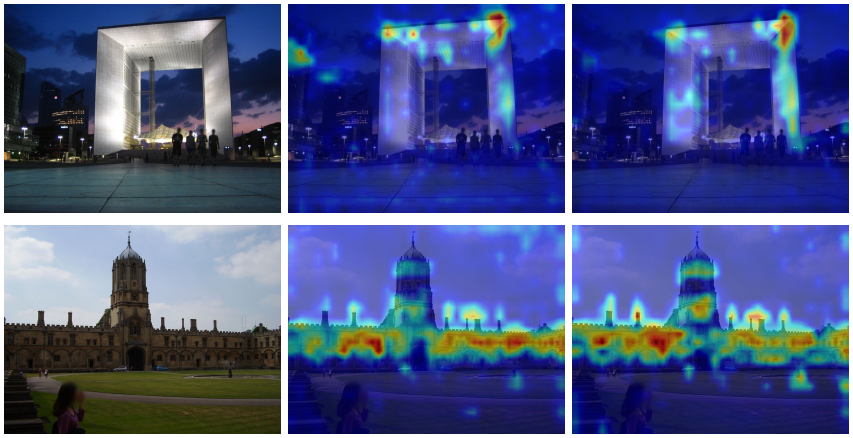
3.2 Spatial Context-Aware Local Attention

In image retrieval tasks, it is essential to train the backbone network to effectively capture fine-grained features. To achieve this, we introduce Spatial Context-Aware Local Attention, designed to capture local context with mid-range interdependencies across two deeper stages, as shown in Fig. 2. Our local self-attention block improves spatial information awareness in key features within a fixed window size, allowing attention calculation between key and query regions, which is then applied to the value. To further enhance the awareness of surrounding context in key features, we employ a Multiscale Context Kernel (MCK) similar to the Contextual Reweighting Network [16] defined as follows:

$$DC_\delta(X)_{i,j} = \sum_{m=-p}^{p} \sum_{n=-p}^{p} w_{m,n} \cdot X_{i+m\cdot\delta, j+n\cdot\delta} + b_{i,j}, \tag{3}$$

$$MCK(X)_{i,j} = \texttt{conv}_{1\times1}\left(\texttt{concat}\left(X, DC_1(X)_{i,j}, DC_2(X)_{i,j}, DC_3(X)_{i,j}\right)\right), \tag{4}$$

where DC, p, $w_{m,n}$, $b_{i,j}$, $\texttt{concat}$, $\texttt{conv}_{1\times1}$ represent the dilated convolution, kernel size of the DC, weight of kernel, positional bias of kernel, channel-wise



(a) Landmark (b) NA (c) SCALA (Ours)

Fig. 3: Visualization of activation maps based on different local self-attention methods.

concatenation, and pointwise convolution. Each convolution kernel is processed through depthwise convolution, and after concatenation, it is computed using pointwise convolution to maintain computational efficiency. After the key of the window size is updated by considering the surrounding local area, the attention score at the pixel (i, j), $A_{i,j}$, is calculated as follows:

$$A_{i,j} = Q_i \cdot MCK(X)^{\top}_{\varepsilon_j(i)} + B_{\varepsilon_j(i)}, \tag{5}$$

$$\mathbf{A}_i^k = \left[A_{i,1}\ A_{i,2}\ \cdots\ A_{i,k}\right]^{\top}, \tag{6}$$

where, $\varepsilon_j(i)$, and $A_{i,j}$ represents the j-th neighbor feature of i, and it's attention score. Note, Q and B denote the projected query feature and the positional bias, respectively. Let $\mathbf{V}_i^k$ be a matrix where each row corresponds to the k-th value projection defined as follows:

$$\mathbf{V}_i^k = \left[V^{\top}_{\varepsilon_1(i)}\ V^{\top}_{\varepsilon_2(i)}\ \cdots\ V^{\top}_{\varepsilon_k(i)}\right]^{\top}. \tag{7}$$

As a result, our local self-attention is defined as follows:

$$\text{SCALA}_k(i) = \texttt{softmax}\left(\frac{\mathbf{A}_i^k}{\sqrt{d}}\right) \cdot \mathbf{V}_i^k, \tag{8}$$

and this process is carried out for each pixel within the feature map. Through our design, we first update the key feature using a convolutional multiscale context kernel, and update the local area by window self-attention based on the updated key values.

SCALA can be interpreted as a context expansion version of Neighborhood Attention (NA) [13]. As shown in Fig. 3, SCALA focuses more activation on the spatial area and reduces unnecessary background attention compared to NA.

3.3 Attention-based Global Feature

After passing through all the layers, the goal is to integrate the feature map into a single global descriptor by considering its universal relationships. This process involves compressing $\mathbf{X} \in \mathbb{R}^{D_4 \times H_4 \times W_4}$, which has important local features activated, into the final descriptor $\mathbf{d} \in \mathbb{R}^{D_4 \times 1}$ using a global pooling method. We have modified the attention-weighted pooling technique previously used in DALG [36] and SimPool [29]. For our attentive pooling, the feature map is projected into key and value features, $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{D_4 \times H_4 \times W_4}$. Then, the attention score is calculated as follows:

$$\mathbf{d}_q = W_q \cdot \left(\frac{1}{H_4 W_4} \sum_{h=1}^{H_4} \sum_{w=1}^{W_4} \mathbf{X}^{\gamma}(h, w) \right)^{\frac{1}{\gamma}}, \tag{9}$$

$$\mathbf{A} = \texttt{softmax}(\frac{\mathbf{K}^{\top} \mathbf{d}_q}{\sqrt{d}}), \tag{10}$$

where $\mathbf{d}_q$, and γ represent the query descriptor, and pooling parameter. First, we obtain the initial global representation through GeM [31] and embed it as a query descriptor through linear projection $\mathbf{d}_q \in \mathbb{R}^{D_4}$. After that, the attention score is calculated by performing matrix multiplication between $\mathbf{K}$ and $\mathbf{d}_q$ followed by applying the softmax function. The calculated attention scores represent the weights for pairwise interactions in the spatial features. Additionally, the value feature is scaled through the following nonlinear elementwise function defined in 11:

$$f_\alpha(x) := \begin{cases} x^{\frac{1-\alpha}{2}}, & \text{if } \alpha \neq 1, \\ \ln x, & \text{if } \alpha = 1. \end{cases} \tag{11}$$

Finally, the scaled value is un-scaled through the inverse function f_α^{-1} after performing matrix multiplication with the attention score, resulting in the final global representation $\mathbf{d} \in \mathbb{R}^{D_4}$ as follows:

$$\mathbf{d} = f_\alpha^{-1}(f_\alpha(\mathbf{V})\mathbf{A}). \tag{12}$$

3.4 Training Objective

We utilize a margin-based softmax loss to train our model on a dataset labeled with landmark classes. For computing the logit values, we used MadaCos [50] as the header. This method dynamically adjusts the scale and margin during the training period, following the equations below:

$$s = \frac{\log\left(\frac{(1-\epsilon)(1-\rho)}{\epsilon\rho}\right)}{1 - \texttt{median}\left(\{\cos(\theta_i)\}_{i=1}^{N}\right)}, \tag{13}$$

$$m = \frac{1}{N}\sum_{i=1}^{N}\cos(\theta_i) - \frac{1}{s}\log\left(\frac{\rho\sum\exp(s\cdot\cos(\theta_f))}{1-\rho}\right), \tag{14}$$

where ρ denotes a hyperparameter for the anchor point of the target probability. The final loss function is defined as follows:

$$\mathcal{L} = -\log\left(\frac{\exp(s\cdot\cos(\theta_t - m))}{\exp(s\cdot\cos(\theta_t - m) + \sum\exp(s\cdot\cos(\theta_f))}\right), \tag{15}$$

where θ_t represents the angle with the label encoding vector of the true class, and θ_f represents the angles with the label encoding vectors of the other classes.

4 Experiments

4.1 Implementation Details

Datasets and Metric We use the Google Landmarks v2-clean dataset (GLDv2-clean) [45] for our training, which is widely used for landmark recognition and retrieval. The GLDv2-clean dataset comprises a total of 1,580,470 images from 81,313 landmarks, featuring a variety of landmarks. To evaluate our model, we primarily use the Oxford and Paris datasets with revisited annotations, referred to as $\mathcal{R}$Oxf, $\mathcal{R}$Par, and +1M distractor datasets [30]. +1M is composed of a large number of challenging images, making it suitable for evaluating the ability to accurately match small similarities in images that depict the same object but appear different overall. The $\mathcal{R}$Oxf, $\mathcal{R}$Par, and +1M datasets contain 4,993, 6,322, and 1,001,001 high-resolution images, respectively. Each dataset has a distinct query set, both comprising 70 images. Image retrieval performance is evaluated based on mean Average Precision (mAP).

Model Architecture Details Our model is built based on the architecture of the Uniformer [20]. While we adopt the architecture of the Uniformer, we completely replace its local and global multi-head relation aggregators [20] with our proposed CDConv and SCALA blocks. Our proposed architecture is similar to the combination of convolution stem and ViT-based backbone in previously proposed hybrid models [35] for instance image retrieval, but it can be considered a lighter and more efficient structure. To implement local self-attention in stages 3 and 4 of our model, we utilized an efficient CUDA extension named $\mathcal{N}$ATTEN [13]. In $\mathcal{N}$ATTEN, we employ a tiled neighborhood attention function to handle the operations between contextually updated keys and queries, as well as the computations between the calculated attention weights and values. We design the proposed model in two versions, small (ULTRON-S) and base (ULTRON-B). For the small model, the number of encoder blocks for stages 1, 2, 3, and 4 are configured as {3, 5, 9, 5}, respectively, and for the base model, they are configured as {5, 7, 18, 5}. In the CDConv, the thresholds τ_1 and τ_2 were set to the top 50% and 75% score of channel attention, and the $\delta_1 = 3$ and $\delta_2 = 6$. Inside SCALA, the kernel size p for DC is set to 3. In both cases, the kernel size for CDConv and the window size for SCALA are set to 7.

Training Details Following previous studies [24, 6, 48], we randomly split the GLDv2-clean dataset, utilizing 80% for our training dataset and the remaining 20% for validation. In the process of image augmentation, random cropping, and color jittering are initially applied, followed by resizing all the enhanced images to 512 $\times$ 512 pixels for use as model inputs. Each blocks are initialized from ImageNet pre-trained weights. We use a batch size of 128 to train our models on four NVIDIA RTX A6000 GPUs. During our training process, we employed the technique of optimizer switching [17]. To elaborate, we trained with the AdamW optimizer for the first 10 epochs, which included a 5-epoch warm-up phase scaling up from a learning rate of $1e{-}3$ to a base learning rate of $1e{-}2$. Subsequently, we used the SGD optimizer with a learning rate of $4e{-}3$ and a momentum of 0.9 for the next 40 epochs. Both methods employed a weight decay of $1e{-}4$. For the learning rate adjustment, we use a cosine learning rate schedule starting after the warm-up epochs and both AdamW and SGD share the same cosine scheduler steps. For the classification loss using MadaCos, ρ is set at 0.04. In our pooling methods, we use learnable parameters γ for initial GeM pooling and set α to 2.0 for scaling.

Inference Details For inference, we perform feature extraction and matching. Query images are cropped according to bounding box coordinates. Following previous works [24, 6, 48], we use an image pyramid at test time to generate multiscale representations for feature extraction. We produce 5 scales: $\left\{\frac{1}{2\sqrt{2}}, \frac{1}{2}, \frac{1}{\sqrt{2}}, 1, \sqrt{2}\right\}$ with 512 dimensions to create compact global descriptors. Each descriptor extracted at multiple scales undergoes L^2 normalization and is then averaged. The averaged descriptor is further refined through L^2 normalization to produce the final descriptor.

4.2 Experimental Results

Comparison with State-of-the-art Methods Table 1 presents a performance comparison on Revisited Oxford ($\mathcal{R}$Oxf) and Revisited Paris ($\mathcal{R}$Par) by incorporating 1 million distractors in the tests (+1M). Our experimental study compares three groups of state-of-the-art methods: (a) refers to two-stage models that perform image retrieval based on global features followed by reranking using local features. (b) refers to models that effectively aggregate local features to generate a single global vector. (c) includes global single-pass models that have recently achieved state-of-the-art performance. All models in (c) learn a single global representation using an angular margin-based classification loss.

When compared to group (a), our proposed ULTRON-S model achieved superior performance across all datasets with single global retrieval, outperforming methods such as geometric verification [6] and reranking transformer [37], without the need for reranking. In comparison with CVNet, ULTRON-B shows lower performance on the $\mathcal{R}$Oxf dataset, but ULTRON-B outperforms CVNet slightly on the $\mathcal{R}$Par dataset. This indicates that our proposed method enables effective searching in large-scale datasets without the need for reranking.

Method		Medium				Hard			
		$\mathcal{R}$Oxf	$\mathcal{R}$Oxf+1M	$\mathcal{R}$Par	$\mathcal{R}$Par+1M	$\mathcal{R}$Oxf	$\mathcal{R}$Oxf+1M	$\mathcal{R}$Par	$\mathcal{R}$Par+1M
(a) Global feature → Local features re-ranking									
R101-DELG+GV [6]	*ECCV20*	78.5	62.7	82.9	62.6	59.3	39.3	65.5	37.0
50-DELG+RRT [37]	*ICCV21*	78.1	67.0	86.7	69.8	60.2	44.1	75.1	49.4
R50-CVNet+CV [19]	*CVPR22*	**87.9**	80.7	90.5	82.4	75.6	65.1	80.2	67.3
R101-CVNet+CV [19]	*CVPR22*	87.2	**81.9**	**91.2**	**83.8**	**75.9**	**67.4**	**81.1**	**69.3**
(b) Global single pass (Local features aggregation)									
R101-HOW-VLAD [41]	*ECCV20*	73.5	60.4	82.3	62.6	52.0	33.2	67.0	41.8
R101-HOW-ASMK [41]	*ECCV20*	80.4	70.2	85.4	68.8	62.5	45.4	70.8	45.4
R50-Token [46]	*AAAI22*	80.5	68.3	87.6	73.9	62.1	43.4	73.8	53.3
R101-Token [46]	*AAAI22*	**82.3**	**70.5**	**89.3**	**76.7**	**66.6**	**47.3**	**78.6**	**55.9**
(c) Global single pass (Global features with classfication loss)									
R101-GeM [45]	*CVPR20*	74.2	-	84.9	-	51.6	-	70.3	-
R101-DELG [6]	*ECCV20*	76.3	63.7	86.6	70.6	55.6	37.5	72.4	46.9
R101-DOLG† [48]	*ICCV21*	78.8	73.1	89.0	79.2	59.4	48.3	73.6	61.5
R50-SENet-$\mathcal{L}_{cls}$ [18]	*CVPR23*	81.4	72.9	90.5	79.0	62.3	48.7	<u>80.3</u>	59.9
R101-SENet-$\mathcal{L}_{cls}$ [18]	*CVPR23*	80.0	72.5	(**91.6**)	82.1	61.7	49.2	(**82.2**)	64.6
R101-SENet-$\mathcal{L}^{\dagger}_{cls}$ [18]	*CVPR23*	81.5	73.4	90.0	80.7	63.4	51.2	78.9	63.8
R50-SpCa [49]	*ICCV23*	79.9	72.8	87.4	78.0	59.3	49.3	73.1	58.3
R101-SpCa [49]	*ICCV23*	<u>83.2</u>	**77.8**	90.6	79.5	65.9	<u>53.3</u>	80.0	65.0
R50-MadaCos† [50]	*ICCV23*	81.8	72.7	90.5	81.4	63.4	50.7	79.9	61.9
R101-MadaCos† [50]	*ICCV23*	**83.5**	73.2	90.1	82.7	<u>66.3</u>	51.4	79.1	64.2
ULTRON-S (Ours)		80.9	71.8	<u>90.8</u>	<u>84.5</u>	61.2	48.8	79.9	<u>68.0</u>
ULTRON-B (Ours)		82.3	<u>73.9</u>	**91.5**	**86.8**	**66.5**	**54.5**	**81.4**	**71.7**

Table 1: **Performance comparison with the previous state-of-the-art models on the standard benchmarks.** +GV, +RRT, and +CV indicate the application of geometric verification, reranking transformer, and correlation verification in the re-ranking process, respectively. † denotes the reproduced result in our training setting. The best and the second-highest performances are indicated in bold and underlined, respectively. Parentheses indicate instances where the official performance exceeded the results reproduced in our training setting.

In group (b), our model demonstrated advanced performance across all benchmark datasets compared to VLAD [1] or ASMK [41] methods by using global features directly extracted from the feature map. When compared to Token [46], our proposed ULTRON-S and ULTRON-B both outperformed the R50- and R101-based Token models, respectively. Particularly, in the comparison between the R101 model and ULTRON-B on the +1M $\mathcal{R}$Paris dataset, our model showed a significant performance improvement of 10.1% and 15.8% in the medium and hard settings, respectively. This can be attributed to the likelihood that the aggregation process of local features in Token leads to a loss of detail, whereas ULTRON's deep global features likely retain their detailed information.

In comparison with group (c), which employs a similar methodology to ours, our proposed approach achieved competitive performance on standard benchmark datasets relative to previous methods. ULTRON-S achieved superior improvements across all datasets compared to DELG [6] and DOLG [48]. This sug-

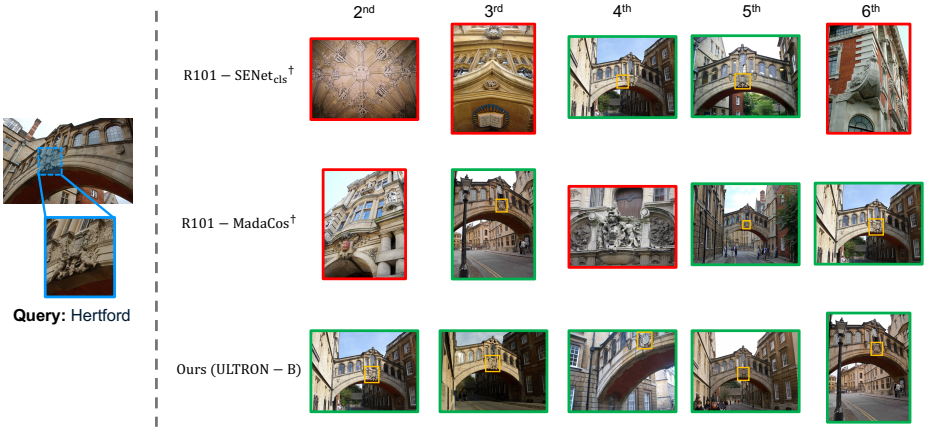


Fig. 4: Highlighted qualitative results with R101-SENet$_{cls}$, R101-MadaCos, and ULTRON-B. Green and Red borders indicate correct and wrong predictions, respectively.

gests that our proposed hybrid network is more effective than ResNet [47] as a backbone for landmark recognition. When benchmarked against recent state-of-the-art models, ULTRON-S showed slightly lesser performance than ResNet50-based SENet [18], SpCa [49], and MadaCos [50] on the $\mathcal{R}$Oxf dataset. However, it achieved comparable or better results on the $\mathcal{R}$Par dataset. This points to a more substantial improvement in performance on the +1M distraction dataset, with ULTRON-S outperforming SENet, SpCa, and MadaCos by 8.1%p, 9.7%p, and 6.1%p, respectively. Among larger models, ULTRON-B demonstrated a decrease in performance compared to the previous state-of-the-art, falling short by 1.2%p and 3.9%p on the $\mathcal{R}$Oxf and $\mathcal{R}$Oxf+1M medium datasets, respectively, and by 0.8%p on the $\mathcal{R}$Par hard dataset. However, it showed similar or improved performance on the remaining datasets. Particularly, ULTRON-B achieved significant performance improvements on the RParis+1M dataset, with increases of 4.1%p (*cf.*ULTRON-B vs. R101-MadaCos†) and 7.9%p (*cf.*ULTRON-B vs. R101-SENet†) in the medium and hard settings, respectively.

Qualitative Results Fig. 4 compares the top-10 retrieval results from the SOTA models and our ULTRON. The SENet [18] results in the first row demonstrate strong performance by robustly embedding inherent structural features based on self-similarity, though they can exhibit weaknesses when handling structurally similar query instances. In comparison with the second and third rows, when trained with the same MadaCos [50] margin loss, our proposed model demonstrates more accurate retrieval performance than ResNet. This indicates that hybrid models such as ULTRON can achieve superior global representation compared to ResNet.

Method	Stage 1-2 (CDConv)	Stage 3-4 (SCALA)	Medium $\mathcal{R}$Oxf	Medium $\mathcal{R}$Par	Hard $\mathcal{R}$Oxf	Hard $\mathcal{R}$Par
baseline [20]			72.5	85.8	51.1	71.0
-	✓		75.1	88.1	52.0	74.5
-		✓	78.7	89.6	58.4	77.8
ULTRON-S(Ours)	✓	✓	**80.3**	**90.7**	**60.7**	**79.5**

Table 2: Effectiveness of our feature embedding blocks in the hybrid network.

Method	FLOPs	Medium $\mathcal{R}$Oxf	Medium $\mathcal{R}$Par	Hard $\mathcal{R}$Oxf	Hard $\mathcal{R}$Par
SA [8]	$3hwd^2 + 2h^2w^2d$	75.1	88.1	52.0	74.5
Swin [21]	$3hwd^2 + 2hwdk^2$	76.5	88.9	55.4	76.1
NA [13]	$3hwd^2 + 2hwdk^2$	78.2	89.7	57.3	77.4
SCALA(Ours)	$3hwd^2 + hwd(2k^2 + 3p^2 + 1)$	**80.3**	**90.7**	**60.7**	**79.5**

Table 3: Performance comparison of self-attention methods with ULTRON-S.

Method	#Param. (M)	FLOPs (G)	Latency (ms)
R101-DOLG [48]	47	334	**110**
R50+ViT-B/16 [35]	101	1096	632
ULTRON-B (Ours)	**43**	**314**	192

Table 4: Computational cost comparison of backbone models with ULTRON-B.

4.3 Ablation Studies

In this section, we empirically demonstrate the superiority of the proposed method through additional experiments. All models use ULTRON-S as the backbone and are trained for 40 epochs.

Table 2 shows the performance changes when our method is applied to the baseline network. Specifically, we replace Uniformer's local multi-head relation aggregator with our proposed channel-wise dilated convolution in stages 1-2 and replace Uniformer's global multi-head relation aggregator with our proposed SCALA blocks in stages 3-4. Applying channel-wise dilated convolution resulted in significant improvements: 2.6%p and 2.3%p on the $\mathcal{R}$Oxf and $\mathcal{R}$Par medium benchmark, and 3.5%p on the $\mathcal{R}$Par hard benchmark. In Stage 3-4, our SCALA-based transformer blocks significantly improved performance, achieving notable gains of 7.3%p and 6.8%p on the $\mathcal{R}$Oxf and $\mathcal{R}$Par hard benchmarks, respectively, compared to Uniformer. These results demonstrate the effectiveness of both CDConv and SCALA quantitatively.

Table 3 presents experimental results comparing our SCALA with previous self-attention mechanisms, demonstrating its superiority. For input feature maps with dimensions $h \times w \times d$, where d denotes the number of channels and h and w represent the height and width of the feature map, respectively. Compared to NA, SCALA has additional FLOPs equivalent to $hwd(3p^2 + 1)$ in order to enhance

the spatial context of the key through the use of MCK. Compared to ViT's self-attention, local self-attention yielded superior results, and our proposed method showed significant performance improvements. On average, it achieved a 3.6%p increase over Swin and a 2.2%p increase over NA, due to the extended spatial context in the key features without substantial increases in computational complexity, allowing for these gains without sacrificing FLOPs.

Table 4 compares the computational cost of representative CNN and ViT-based models in landmark image retrieval with our model. The evaluation was carried out with a 1024×1024 feature map as input, using 5 scaling for embedding. Note that for [35], since the official code is not yet publicly available, the actual computational cost is likely higher than reported, as only the backbone's cost was measured. Our method has fewer learnable parameters and FLOPs compared to previous backbone models. In particular, when compared to the backbone used in [35], it shows significantly lower computational cost and approximately $\times 3$ times faster feature extraction. This is attributed to the efficient design of our backbone, which mitigates the necessity for supplementary modules to embed detailed information by enhancing local features. However, due to the limitations of the local self-attention implementation, it is slower than CNN-based models, despite being faster than universal self-attention models. This indicates significant potential for improvement, especially with enhancements in CUDA APIs supporting window self-attention.

5 Conclusion

We introduced ULTRON, a novel hybrid model for large-scale landmark recognition and retrieval. ULTRON addresses ViT's limitations in fine-grained feature capture through SCALA, enhancing spatial context awareness while reducing unnecessary background attention. Additionally, CDConv adjusts dilation rates based on channel importance, improving both local and global feature learning. This multiscale approach reduces redundancy and provides robust feature embedding. ULTRON outperformed state-of-the-art models on challenging benchmarks, such as the one million distractors. For future works, we aim to implement CDConv with continuous dilation sizes for enhanced efficiency and improve the latency of local self-attention to develop a convolution-free local transformer for landmark recognition.

Acknowledgments. This work was supported in part by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2024-00358935, 50%), in part by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00222799, 20%), in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2024-00344883, 20%), and in part by the Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2024-RS-2023-00254177, 10%) grant funded by the Korea government(MSIT).

References

1. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: CVPR. pp. 5297–5307 (2016)
2. Avrithis, Y., Tolias, G.: Hough pyramid matching: Speeded-up geometry re-ranking for large scale image retrieval. IJCV **107**, 1–19 (2014)
3. Babenko, A., Lempitsky, V.: Aggregating local deep features for image retrieval. In: ICCV. pp. 1269–1277 (2015)
4. Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V.: Neural codes for image retrieval. In: ECCV. pp. 584–599. Springer (2014)
5. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: ECCV. pp. 404–417. Springer (2006)
6. Cao, B., Araujo, A., Sim, J.: Unifying deep local and global features for image search. In: ECCV. pp. 726–743. Springer (2020)
7. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 **5** (2017)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. El-Nouby, A., Neverova, N., Laptev, I., Jégou, H.: Training vision transformers for image retrieval. arXiv preprint arXiv:2102.05644 (2021)
10. Fischler, M.A., Bolles, R.C.: 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)
11. Gordo, A., Almazan, J., Revaud, J., Larlus, D.: End-to-end learning of deep visual representations for image retrieval. International Journal of Computer Vision **124**(2), 237–254 (2017)
12. Gordo, A., Radenovic, F., Berg, T.: Attention-based query expansion learning. In: ECCV. pp. 172–188. Springer (2020)
13. Hassani, A., Walton, S., Li, J., Li, S., Shi, H.: Neighborhood attention transformer. In: CVPR. pp. 6185–6194 (2023)
14. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
15. Jgou, H., Perronnin, F., Douze, M., Snchez, J., Prez, P., Schmid, C.: Aggregating local image descriptors into compact codes. IEEE TPAMI **34**(9), 1704–1716 (2012)
16. Jin Kim, H., Dunn, E., Frahm, J.M.: Learned contextual feature reweighting for image geo-localization. In: CVPR. pp. 2136–2145 (2017)
17. Keskar, N.S., Socher, R.: Improving generalization performance by switching from adam to sgd. arXiv preprint arXiv:1712.07628 (2017)
18. Lee, S., Lee, S., Seong, H., Kim, E.: Revisiting self-similarity: Structural embedding for image retrieval. In: CVPR. pp. 23412–23421 (2023)
19. Lee, S., Seong, H., Lee, S., Kim, E.: Correlation verification for image retrieval. In: CVPR. pp. 5374–5384 (2022)
20. Li, K., Wang, Y., Zhang, J., Gao, P., Song, G., Liu, Y., Li, H., Qiao, Y.: Uniformer: Unifying convolution and self-attention for visual recognition. IEEE TPAMI (2023)
21. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 10012–10022 (2021)

22. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60**, 91–110 (2004)
23. Ng, T., Balntas, V., Tian, Y., Mikolajczyk, K.: Solar: second-order loss and attention for image retrieval. In: ECCV. pp. 253–270. Springer (2020)
24. Noh, H., Araujo, A., Sim, J., Weyand, T., Han, B.: Large-scale image retrieval with attentive deep local features. In: ICCV. pp. 3456–3465 (2017)
25. Perronnin, F., Liu, Y., Sánchez, J., Poirier, H.: Large-scale image retrieval with compressed fisher vectors. In: CVPR. pp. 3384–3391. IEEE (2010)
26. Phan, L., Nguyen, H.T.H., Warrier, H., Gupta, Y.: Patch embedding as local features: Unifying deep local and global features via vision transformer for image retrieval. In: ACCV. pp. 2527–2544 (2022)
27. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: CVPR. pp. 1–8. IEEE (2007)
28. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In: CVPR. pp. 1–8. IEEE (2008)
29. Psomas, B., Kakogeorgiou, I., Karantzalos, K., Avrithis, Y.: Keep it simpool: Who said supervised transformers suffer from attention deficit? In: ICCV. pp. 5350–5360 (2023)
30. Radenović, F., Iscen, A., Tolias, G., Avrithis, Y., Chum, O.: Revisiting oxford and paris: Large-scale image retrieval benchmarking. In: CVPR. pp. 5706–5715 (2018)
31. Radenović, F., Tolias, G., Chum, O.: Fine-tuning cnn image retrieval with no human annotation. IEEE TPAMI **41**(7), 1655–1668 (2018)
32. Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., Shlens, J.: Stand-alone self-attention in vision models. Advances in neural information processing systems **32** (2019)
33. Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: Cnn features off-the-shelf: an astounding baseline for recognition. In: CVPRW. pp. 806–813 (2014)
34. Song, C.H., Han, H.J., Avrithis, Y.: All the attention you need: Global-local, spatial-channel attention for image retrieval. In: WACV. pp. 2754–2763 (2022)
35. Song, C.H., Yoon, J., Choi, S., Avrithis, Y.: Boosting vision transformers for image retrieval. In: WACV. pp. 107–117 (2023)
36. Song, Y., Zhu, R., Yang, M., He, D.: Dalg: Deep attentive local and global modeling for image retrieval. arXiv preprint arXiv:2207.00287 (2022)
37. Tan, F., Yuan, J., Ordonez, V.: Instance-level image retrieval using reranking transformers. In: ICCV. pp. 12105–12115 (2021)
38. Teichmann, M., Araujo, A., Zhu, M., Sim, J.: Detect-to-retrieve: Efficient regional aggregation for image search. In: CVPR. pp. 5109–5118 (2019)
39. Tolias, G., Avrithis, Y., Jégou, H.: To aggregate or not to aggregate: Selective match kernels for image search. In: ICCV. pp. 1401–1408 (2013)
40. Tolias, G., Avrithis, Y., Jégou, H.: Image search with selective match kernels: aggregation across single and multiple images. IJCV **116**, 247–261 (2016)
41. Tolias, G., Jenicek, T., Chum, O.: Learning and aggregating deep local descriptors for instance-level recognition. In: ECCV. pp. 460–477. Springer (2020)
42. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of cnn activations. arXiv preprint arXiv:1511.05879 (2015)
43. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: CVPR. pp. 11534–11542 (2020)
44. Wengert, C., Douze, M., Jégou, H.: Bag-of-colors for improved image search. In: ACM MM. pp. 1437–1440 (2011)

45. Weyand, T., Araujo, A., Cao, B., Sim, J.: Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In: CVPR. pp. 2575–2584 (2020)
46. Wu, H., Wang, M., Zhou, W., Hu, Y., Li, H.: Learning token-based representation for image retrieval. In: AAAI. vol. 36, pp. 2703–2711 (2022)
47. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: CVPR. pp. 1492–1500 (2017)
48. Yang, M., He, D., Fan, M., Shi, B., Xue, X., Li, F., Ding, E., Huang, J.: Dolg: Single-stage image retrieval with deep orthogonal fusion of local and global features. In: ICCV. pp. 11772–11781 (2021)
49. Zhang, Z., Wang, L., Zhou, L., Koniusz, P.: Learning spatial-context-aware global visual feature representation for instance image retrieval. In: ICCV. pp. 11250–11259 (2023)
50. Zhu, Y., Gao, X., Ke, B., Qiao, R., Sun, X.: Coarse-to-fine: Learning compact discriminative representation for single-stage image retrieval. In: ICCV. pp. 11260–11269 (2023)