Supplementary Material: ULTRON

Minseong Kweon[1] and Jinsun Park[2,3⋆]

[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 this supplementary material, we provide the motivation for our approach and in-depth explanations of our proposed method with detailed pseudo-code. Additionally, we include extensive qualitative results with visualizations.

1 Details of the Proposed Method

1.1 Channel-wise Dilated Convolution

This section provides the motivation and detailed implementation of our proposed convolutional encoder, referred to as **C**hannel-wise **D**ilated **Conv**olution (**CDConv**). We aimed to increase the multiscale capacity to enhance fine-grained feature recognition while following the MobileNet [4] blocks for our first and second encoder blocks. Previous studies have suggested methods that combine convolution layers with various dilation rates to increase the multiscale capacity [1, 13, 5]. However, since shallow layers primarily handle low-level features such as edges and textures, which do not require the complex processing capabilities of a heavy encoder, this can result in unnecessary redundancy from the perspective of the feature map at the initial stages of the network.

To provide multiscale awareness while reducing unnecessary redundancy, we design a method that applies different dilation rates to each channel for convolution, rather than concatenating the outputs of multiple convolution layers with various dilation rates in the shallow layers. Determining the criteria for applying different dilation rates to each channel becomes crucial in this context. The importance of a specific channel indicates the significance of the features it represents. Generally, a narrower receptive field is more effective for capturing detailed information in important channels, while a wider receptive field is more efficient for capturing global features in less important channels. Therefore, we propose adjusting the dilation rates based on the importance of each channel: increasing the dilation for less important channels to capture more spatial information and decreasing it for more important channels to capture more regional information. Specifically, we use the Efficient Channel Attention (ECA)

⋆ Corresponding author.

module [11] to compute channel attention. We then segment the channels based on the magnitude of the attention values, assigning smaller dilations to channels with higher attention values. This method is implemented as described in Algorithm 1.

Algorithm 1 Channel-wise Dilated Convolution

```
Input  : F ∈ ℝ^{B×C×H×W}
Output: F̃ ∈ ℝ^{B×C×H×W}
B, C, H, W ← shape(F)
F̃ ← zeros_like(F)
G ← GAP(F)                                    ▷ Global Average Pooling
a ← AdaptiveKernel(G)                         ▷ Adaptive selection of kernel size
a ← sigmoid(a)                                ▷ Sigmoid to get attention score
for c ← 1 to C do
    if a[c] > τ_1 then
        d_c ← 1
    else
        if τ_1 ≥ a[c] > τ_2 then
            d_c ← δ_1
        else
            d_c ← δ_2
        end
    end
end
for b ← 1 to B do
    for c ← 1 to C do
        F̃[b, c, :, :] ← DC(F[b, c, :, :], d_c)
    end
end
return F̃
```

The conventional approach of concatenating results from multiple convolutions typically involves linear projection to adjust dimensions and resolution. In contrast, our method applies convolutions to features masked based on dilation criteria and then simply sums the results, eliminating the need for a linear layer and reducing the computational cost. This approach, equivalent to performing a single convolution operation from the feature map's perspective, reduces unnecessary redundancy. Consequently, CDConv was used to replace the depthwise convolution in MobileNet blocks.

Previous studies like SCA-CNN [2], CBAM [12], and GLAM [9] have utilized both channel and spatial attention for image understanding. Our method differs by adjusting local context enhancement based on channel importance derived from channel attention, allowing for more precise and efficient spatial feature capture. We empirically demonstrate our approach's effectiveness with improved performance. However, further research and continuous implementation of channel-based dilation adjustments, rather than discrete ones, could uncover additional enhancement opportunities yet to be explored.

1.2 Spatial Context-Aware Local Attention

In this section, we describe the motivation and detailed implementation of our proposed local self-attention mechanism, **S**patial **C**ontext-**A**ware **L**ocal **A**ttention (**SCALA**). Recent high-performing ViT-based models [8, 10] have all utilized atrous convolution to enhance the quality of local features before embedding them into a global vector. In contrast, our approach departs from this method by employing window attention to perform self-attention within a limited region. This allows for a better understanding of the context within regional areas.

Algorithm 2 Spatial Context-Aware Local Attention

Input : $F \in \mathbb{R}^{B\times C\times H\times W}$
Output: $\tilde{F} \in \mathbb{R}^{B\times C\times H\times W}$
$B, C, H, W \leftarrow$ `shape`(F)
$\tilde{F} \leftarrow$ `zeros_like`(F)
Function `MCK`(X):
  $X_1 \leftarrow$ `DC`$(X, 1)$ ▷ Convolution with 3 by 3 region
  $X_2 \leftarrow$ `DC`$(X, 2)$ ▷ Dilated convolution with 5 by 5 region
  $X_3 \leftarrow$ `DC`$(X, 3)$ ▷ Dilated convolution with 7 by 7 region
  $X \leftarrow$ `Proj`(`concat`(X, X_1, X_2, X_3))
  return X
Function `SCALA`(F):
  $Q, K, V \leftarrow$ `WindowSplit(Proj`(F)`)` ▷ Reshape feature for local attention
  $Q \leftarrow$ `Reshape(MCK(Reshape`(Q)`)))` ▷ Apply Multiscale Context Kernel
  $\mathrm{a} \leftarrow$ `TiledMatMul`(Q, K) ▷ Mat-mul operation in fixed window size
  $\mathrm{a} \leftarrow \mathrm{a} + \text{relative bias}$ ▷ Put positional bias
  $\mathrm{a} \leftarrow$ `softmax`(a) ▷ Softmax to get attention score
  $\mathrm{a} \leftarrow$ `Dropout`$(\mathrm{a}, \text{attention dropout rate})$
  $F \leftarrow$ `TiledMatMul`(a, V) ▷ Get attentive feature
  $F \leftarrow$ `Reshape(Proj`(F)`)`
  $F \leftarrow$ `Dropout`$(F, \text{proj dropout rate})$
  return F
for $b \leftarrow 1$ **to** B **do**
  for $c \leftarrow 1$ **to** C **do**
    $shortcut \leftarrow F[b, c, :, :]$
    $F[b, c, :, :] \leftarrow$ `Norm`$(F[b, c, :, :])$
    $\tilde{F}[b, c, :, :] \leftarrow$ `SCALA`$(F[b, c, :, :])$ ▷ Operated by CUDA extension
    $\tilde{F}[b, c, :, :] \leftarrow shortcut + \tilde{F}[b, c, :, :]$
    $\tilde{F}[b, c, :, :] \leftarrow \tilde{F}[b, c, :, :] +$ `MLP`(`Norm`$(\tilde{F}[b, c, :, :])$)
  end
end
return $\tilde{F}$

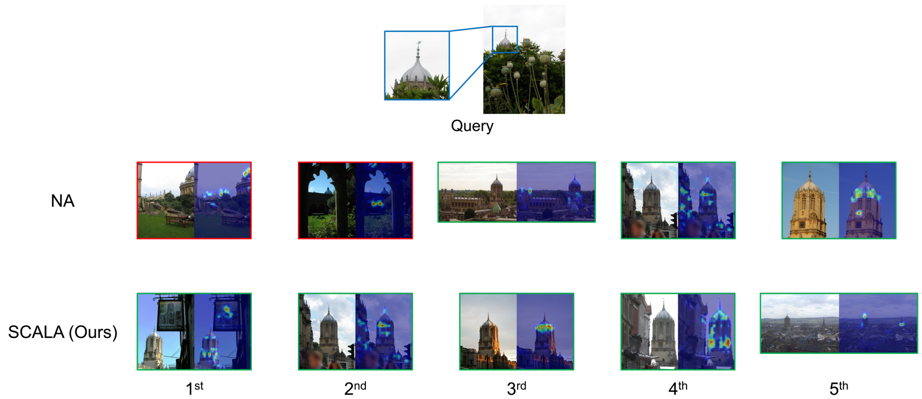

Fig. 1: Qualitative results with different local self-attention: For each image, the classification activation map is visualized together.

We followed the previous method [3] by performing attention updates within fixed regions. However, unlike prior work, we enhance the spatial context awareness of key features. This approach allows us to consider a broader range than the window area where the attention weight is applied to the value. To achieve spatial feature enhancement for the key, we designed a lightweight version of atrous convolution, called the Multiscale Context Kernel (MCK). This was implemented similarly to the method used in the encoder of CRN [5]. The specific operation is detailed in the following Algorithm 2.

Our approach was implemented using the tiled self-attention method from the CUDA extension $\mathcal{N}$ATTEN, released by NA [3], to perform the key-query operation and the weight-value operation. This method utilizes CUDA to allocate and use shared memory, enabling matrix multiplication within the window size. Tiled local self-attention operates by first dividing the input data into small, fixed-size tiles. This allows each thread to read adjacent data cells from global memory and store them efficiently in shared memory. Within each tile, the key-query-value computations are then performed in parallel, leveraging the speed of shared memory access. The attention scores are computed and subsequently normalized using the softmax function. Finally, these normalized attention scores are employed to compute the weighted sum of the value tiles, resulting in the final output. This method significantly enhances computational efficiency and memory usage, facilitating faster and more scalable attention mechanisms.

Fig. 1 visualizes the final classification activations of each network when using the baseline NA and our implemented SCALA method. It can be observed that the proposed method captures more critical points, resulting in improved Top-5 retrieval performance.

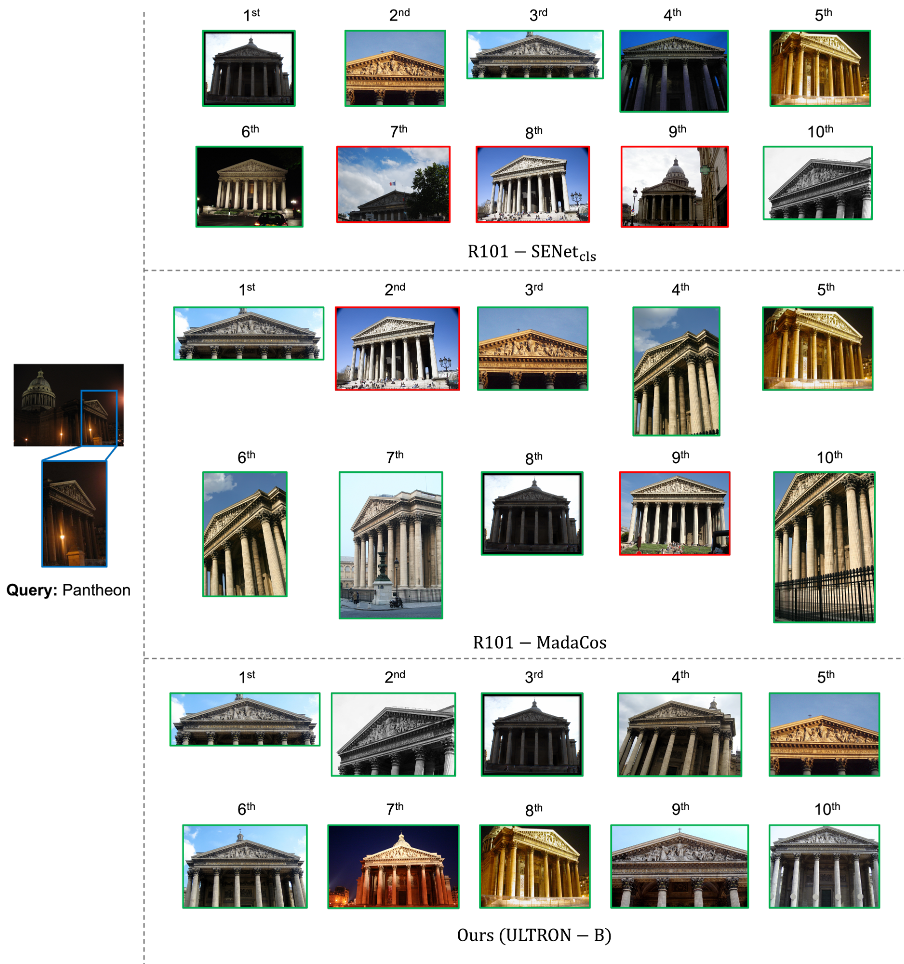

Fig. 2: Top-10 retrieval results for Pantheon with R101-SENet$_{cls}$, R101-MadaCos, and ULTRON-B.

2 Additional Retrieval Results Analysis

We show qualitative results for additional challenging queries, comparing our proposed model with the previous state-of-the-art models [6, 14].

In Fig. 2, the Top-10 retrieval results for the night-time landmark query using the models R101-SENet$_{cls}$, R101-MadaCos, and ULTRON-B are presented. These results illustrate examples where even recent state-of-the-art models, such as SENet and MadaCos, struggle to perform robustly with night-time images. Recently, a GAN-based synthetic-image generator [7] was proposed to improve retrieval performance for night-time query images by converting day-time images into night-time images for training. Our proposed model demonstrates excellent

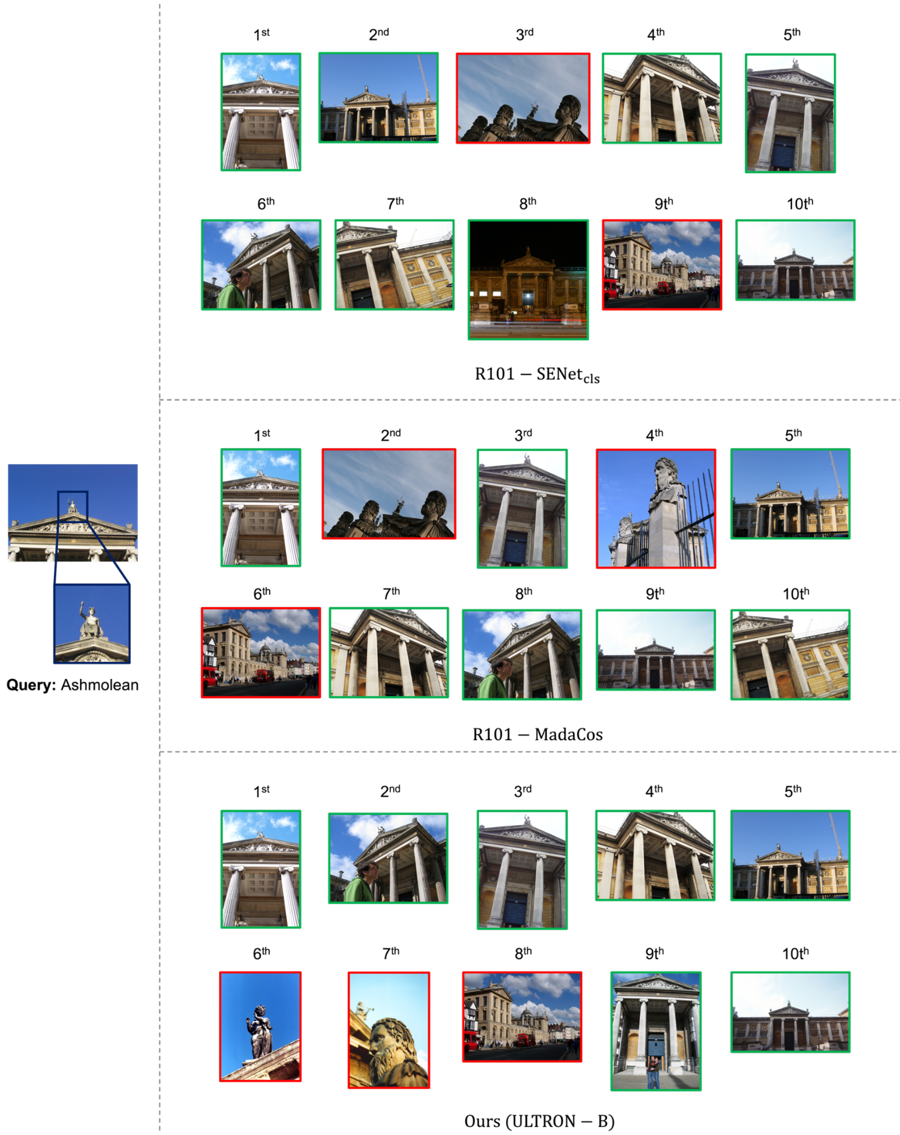

Fig. 3: Top-10 retrieval results for Ashmolean with R101-SENet$_{cls}$, R101-MadaCos, and ULTRON-B.

performance on night-time images without requiring additional methods, achieving no errors in the Top-10 samples.

In Fig. 3, we compare the Top-10 retrieval results of our proposed model with those of previous state-of-the-art models for challenging queries, where error samples and correct samples are visually, structurally, and contextually similar.

When comparing Top-10 performance for this query, SENet demonstrated the highest performance, while our proposed and R101-MadaCos models performed equally. However, unlike the third result of SENet and the second result of MadaCos, which both included errors in the Top-5 results, our proposed model did not exhibit such errors. Furthermore, based on the Top-5 results, our proposed model demonstrated the best performance. These findings highlight the strength and precision of our model in identifying the most relevant images, even when the queries are difficult.

References

1. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 **5** (2017)
2. Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.S.: Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5659–5667 (2017)
3. Hassani, A., Walton, S., Li, J., Li, S., Shi, H.: Neighborhood attention transformer. In: CVPR. pp. 6185–6194 (2023)
4. 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)
5. Jin Kim, H., Dunn, E., Frahm, J.M.: Learned contextual feature reweighting for image geo-localization. In: CVPR. pp. 2136–2145 (2017)
6. Lee, S., Lee, S., Seong, H., Kim, E.: Revisiting self-similarity: Structural embedding for image retrieval. In: CVPR. pp. 23412–23421 (2023)
7. Mohwald, A., Jenicek, T., Chum, O.: Dark side augmentation: Generating diverse night examples for metric learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11153–11163 (2023)
8. 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)
9. 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)
10. Song, C.H., Yoon, J., Choi, S., Avrithis, Y.: Boosting vision transformers for image retrieval. In: WACV. pp. 107–117 (2023)
11. 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)
12. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
13. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
14. 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)