Patch Embedding as Local Features: Unifying Deep Local and Global Features Via Vision Transformer for Image Retrieval

Lam Phan[1], Hiep Thi Hong Nguyen[1], Harikrishna Warrier[1], and Yogesh Gupta[1]

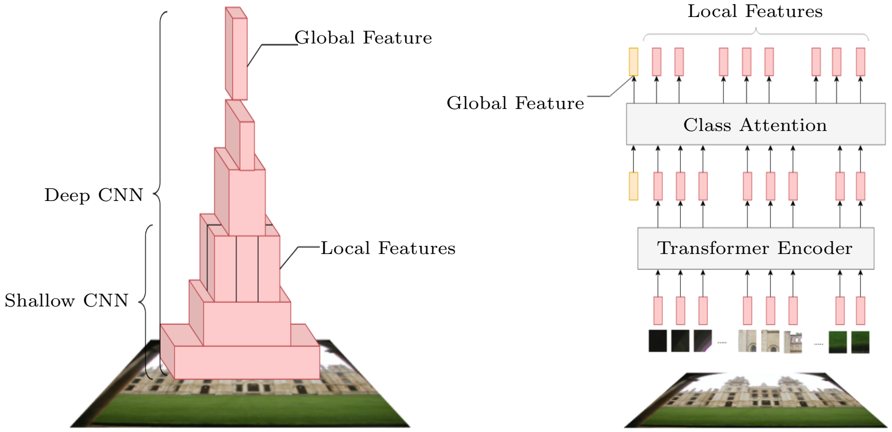
HCL Technologies
lam.phan,hiep.nguyen,harikrishna.w,yogeshg@hcl.com

Abstract. Image retrieval is the task of finding all images in the database that are similar to a query image. Two types of image representations have been studied to address this task: global and local image features. Those features can be extracted separately or jointly in a single model. State-of-the-art methods usually learn them with Convolutional Neural Networks (CNNs) and perform retrieval with multi-scale image representation. This paper's main contribution is to unify global and local features with Vision Transformers (ViTs) and multi-atrous convolutions for high-performing retrieval. We refer to the new model as ViTGaL, standing for Vision Transformer based Global and Local features (ViTGaL). Specifically, we add a multi-atrous convolution to the output of the transformer encoder layer of ViTs to simulate the image pyramid used in standard image retrieval algorithms. We use class attention to aggregate the token embeddings output from the multi-atrous layer to get both global and local features. The entire network can be learned end-to-end, requiring only image-level labels. Extensive experiments show the proposed method outperforms the state-of-the-art methods on the Revisited Oxford and Paris datasets. Our code is available at here

1 Introduction

Image retrieval is an important and long-standing task in computer vision that aims to effectively retrieve all images matching a query image over an (usually very large) image collection. This task is challenging due to various conditions, such as extreme viewpoint/pose, illumination change, occlusion, etc., especially on large-scale datasets. Therefore, image representations that are discriminative enough to deal with these challenges play a central role in this task. There are two types of image representations: global and local features.

Before deep learning revolutionized the field, various handcrafted features [7,28,33,24] have been proposed. With the introduction of deep learning to computer vision, both global feature [2,38,19,48,41] and local features [13,31,30,34,47] are extracted using deep neural network (DNN) in a data-driven paradigm. The global feature summarizes an image, usually as a high-dimensional vector. Due to its compact representation, the global feature can be learned so that it is invariant to viewpoint and illumination with the risk of losing information about the spatial arrangement of visual elements. On the other hand, local features encode detailed spatial features and well preserve geometrical information about specific image regions. They are useful for patch-level matching between images and are shown to be essential for high retrieval precision [9,47]. Therefore, the best retrieval methods [9,44] typically use a global feature to first search for a list of candidate matching images, then re-rank them using local features matching. Recently, [55] proposed to integrate local features and a global feature into a compact descriptor, then perform retrieval in a single stage and showing promising results compared to two stages method.



(a) CNN features extraction. (b) ViT features extraction.

Fig. 1: Global and local features extraction pipeline of CNN and ViT

Today, state-of-the-art methods mainly utilize convolutional neural networks (CNNs) to extract local and global features [9,47,34]. Due to the hierarchical representations of CNNs, global features are associated with deep layers representing high-level cues, while local features are extracted at a shallow layer with high spatial dimensions to preserve spatial information. Recently, [15] demonstrated that Vision Transformers (ViT) are capable of equal or superior performance on image classification tasks at a large scale. Later works also show promising results of these models for other vision tasks, including object detection [27,26,57], semantic segmentation [52,53], video understanding [3,8], monocular depth estimation [40,25], to name a few. Unlike CNN, ViT first divides an image into non-overlapping patches and uses self-attention to aggregate information between these patches, making the spatial dimensions of ViT fixed across layers. Moreover, a recent study [39] found that spatial information from the input is preserved in ViT even as the final layer. In contrast, the representation from deep layers of CNNs is much less spatially discriminative. This property inspires us to utilize the patch embeddings of ViT and treat them as local features to perform geometric verification. Figure 1 shows local and global features extraction pipeline of CNN and ViT. With the ability to preserve spatial information across layers, patch embedding from arbitrary layers of the ViT model can be used as local features. Our experiments show that using patch embeddings from the final layer of ViT yields the best result. Compared to ViT, local features in CNN are extracted at shallow layers, thus may limit its capacity to extract discriminative and robust local features.

One major drawback of the vanilla ViT model is the time and memory complexity of the self-attention operation, which increases quadratically with the number of input patches. For fine-grained retrieval tasks such as landmark retrieval, the best practice is to train retrieval models with large input resolution [9,55,51,47,34], where the largest resolution can be up to 1024×1024 [9,47] during inference. These large resolutions make the direct usage of vanilla ViT models for fine-grained retrieval tasks infeasible. To alleviate the complexity of vanilla ViT, various strategies have been proposed. The dominant approach is reducing the spatial dimension of input resolutions at every block of layers, similar to CNN [27,26,50]. Recently, XCiT [1] replaced a self-attention between tokens with a "transposed" attention between channels which they call "cross-covariance attention" (XCA). Thanks to XCA operation, XCiT reduces the time and memory complexity of the vanilla ViT from quadratically to linear without sacrificing the accuracy of the classification task. Moreover, similar to vanilla ViT, the spatial dimension of input patches remains fixed across layers in XCiT, making it feasible to extract patch embeddings at deep layers. We, therefore, utilize XCiT as the backbone for our retrieval model.

To achieve high retrieval performance, state-of-the-art methods usually utilize an image pyramid at inference time to produce multi-scale representations, thus increasing time and memory complexity. To solve this problem, we proposed to simulate an image pyramid with multi-atrous convolutions [10]. Extensive experiments on Revisited Oxford and Pairs [37] show the effectiveness of the multi-atrous, as we observe significant performance improvements compared to the baseline. Moreover, since XCiT uses XCA operation to approximate the self-attention operation implicitly, local patch interaction implemented by small convolutional layers at each block is added. Therefore, the multi-atrous convolutions could help increase the receptive field of the model, leading to more robust local and global features. To summarize, our main contributions are as follows:

- We propose using the ViT model to perform image retrieval in a two-stage paradigm by using the patch embeddings and treating them as local features to perform geometric verification. Concretely, we use XCiT as our backbone to alleviate the complexity of vanilla ViT.
- We add multi-atrous convolutions to simulate the image pyramid used in standard retrieval algorithms, leading to state-of-the-art results using only single-scale representation.

- Extensive experiments are conducted, and comprehensive analyses are provided to validate the effectiveness of our solution. Our ViT-based model, referred to as ViTGaL (Vision Transformer based Global and Local features), significantly outperforms the previous state-of-the-art CNNs models.

2 Related Work

2.1 Local features

Hand-crafted local features [7,28] based on low-level visual information were widely used in earlier retrieval works. To compare two images with local features, aggregation methods such as [22,58] are usually used. To improve precision and produce reliable and interpretable scores, a second reranking stage based on geometric verification via matching local features with RANSAC [18] is also widely adopted [35,4]. More recently, many methods have been proposed to learn local features [6,29,16,5,14,56,42,34,47] with deep neural networks. Those methods rely on CNN to perform local features extraction, where shallow layers from a CNN backbone are utilized. We go beyond the normal approach of using CNN and propose to use ViT patch embeddings as local features in traditional CNN-based methods. To the best of our knowledge, no previous works have studied whether patch embeddings of ViT can be utilized to perform geometric verification and how they perform compared to CNN.

2.2 Global feature

Before deep learning revolutionalized the fields, conventional approaches to obtaining global features were via aggregating local features [46,22,24]. In deep learning era, most high-performing global features are extracted using neural networks, where the differentiable version of the traditional aggregating method [38,2,54] is used to enable end-to-end training which either ranking-based loss [41,11,20] or classification loss [49,12]. Unlike those widely used for the CNN-based model, our work uses multi-head self-attention operation at classification token to obtain global descriptor. For a fair comparison to previous state-of-the-art CNNs [9,55], we used ArcFace loss [12] to train our retrieval model.

2.3 Joint local and global CNN features

Using both global and local features for retrieval is shown to be more efficient than using either global or local features alone [9,55]. Therefore, it is natural to consider learning both features jointly since using separate models may lead to high memory usage and increased latency. [43] distills pre-trained local and global features into a single model. DELG [9] takes a step further and proposes to train local and global features in an end-to-end manner jointly. We follow the work of DELG and also present a unifying model. The difference is our model is ViT-based, while DELG and other conventional models are CNN-based.

2.4 Transformers for high-resolution images

Since fine-grained retrieval tasks such as landmark retrieval require high-resolution input to achieve high retrieval performance, [9,55], using vanilla ViT for those tasks is infeasible. Recently, several works have adopted visual transformers for high-resolution images and proposed multiple strategies to alleviate the complexity of vanilla ViT. [50] designed a pyramidal architecture and addresses complexity by gradually reducing the spatial resolution of keys and values. Since then, several works have adopted the idea of lowering spatial resolution at each layer for efficient computations. [17] utilized pooling to reduce the resolution across the spatial and temporal dimensions, while [27] used local attention with shifted windows and patch merging. Recently, XCiT [1] proposed to replace the quadratic self-attention operation with a "transposed" attention operation between channels which they call "cross-covariance attention" (XCA). The advantage of XCiT is that it preserves the spatial dimension across layers, making it feasible to extract patch embeddings at deep layers. We, therefore, utilize XCiT as the backbone for our retrieval model.

3 Methodology

3.1 Model

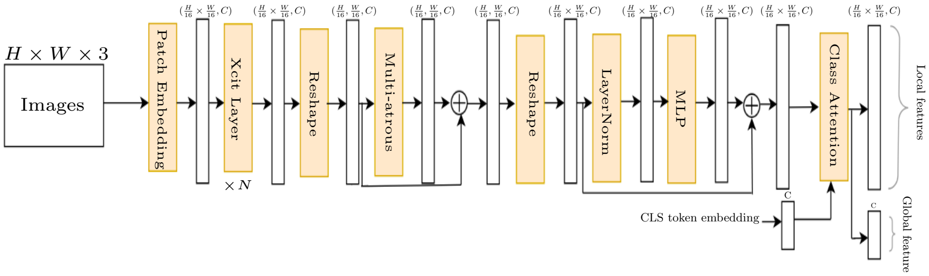

$H \times W \times 3$
Images
Patch Embedding
$(\frac{H}{16} \times \frac{W}{16}, C)$
Xcit Layer
$\times N$
$(\frac{H}{16} \times \frac{W}{16}, C)$
Reshape
$(\frac{H}{16}, \frac{W}{16}, C)$
Multi-atrous
$(\frac{H}{16}, \frac{W}{16}, C)$
$(\frac{H}{16}, \frac{W}{16}, C)$
Reshape
$(\frac{H}{16} \times \frac{W}{16}, C)$
LayerNorm
$(\frac{H}{16} \times \frac{W}{16}, C)$
MLP
$(\frac{H}{16} \times \frac{W}{16}, C)$
$(\frac{H}{16} \times \frac{W}{16}, C)$
Class Attention
$(\frac{H}{16} \times \frac{W}{16}, C)$
CLS token embedding
C
C
Local features
Global feature

Fig. 2: The architecture of our ViTGal model. It consists of a ViT-based model XCiT [1] as a backbone; a multi-atrous convolution followed by a LayerNorm and an MLP layer to simulate an image pyramid, and a class attention layer to aggregate the token embeddings into a compact representation for global retrieval as well as new tokens embeddings for geometric verification.

Our VitGaL model is depicted in Figure 2. We propose utilizing patch embeddings at the final layer of the ViTGaL model for geometric verification in the reranking stage. We also merge all the attention scores in different attention heads in the class attention layer and extract associated patch embeddings with the top scores.

Given an image $I \in R^{H \times W \times 3}$, we follow [1] and reshape the image into a sequence of flattened 2D patches $x_p \in R^{N \times (P^2.3)}$, where (H, W) is the resolution of the original image, (P, P) is the resolution of each image patch, and $N = H \times W/P^2$ is the resulting number of patches. These patches first undergo the patch embedding layer to transform into tokens of dimension D, which we follow [1] and implement by a small CNN. Those tokens are then added with the positional embeddings and fed into N XCiT layer, which we also implement following [1].

To simulate an image pyramid used in multiple state-of-the-art methods, we propose to use multi-atrous convolutions, as depicted in figure 2. We first reshape the 1d tokens sequence into a tensor of shape $H/P \times W/P \times C$. This tensor then goes to the multi-atrous layer and a skip connection to propagate information from the XCiT layer. The multi-atrous module contains five dilated convolution layers to obtain feature maps with different spatial receptive fields and a global average pooling branch. To save the computational and memory of our model, each dilated convolution have the output channel dimension of 1/6 of the input channel dimension. These features are concatenated and processed by a depthwise convolution layer. We then reshape the 3d tensor back to the 1d sequence. Finally, we add layer norm, MLP, and skip connection layer following the design of the FFN module used in the self-attention layer in ViT.

To aggregate the token embeddings into a compact representation, we add a cls token and use a single class attention layer. This layer is identical to the transformer encoder block used in ViT, except the self-attention operation is only calculated between the cls token embedding (treated as a query) and the token embeddings of image patches (treated as keys and values). The tokens embeddings output from the class attention layer will be used for the reranking stage with geometric verification, while the cls token embedding is used as the global feature.

Furthermore, since hundreds to thousands of local features are used in the reranking stage, they must be represented compactly. Moreover, [23] shows that whitening down weights co-occurrences of local features, which is generally beneficial for retrieval applications. We, therefore, implement our local features dimensionality reduction using a small autoencoder (AE) module [21] following the state-of-the-art dimensionality reduction method used in [9], as depicted in figure 3. Both encoder and decoder are implemented by a simple multilayer perceptron (MLP), where we set the number of layers to 1. The new local features are obtained as $\mathcal{L} = \mathrm{E}(\mathcal{S})$, where $\mathcal{L} \in (H/16 \times W/16, C_E)$, $\mathcal{S} \in (H/16 \times W/16, C)$ is the local features from the trained ViTGaL model and E is the encoding part of the autoencoder. Note that the parameters of ViTGaL are kept fixed during the training of the autoencoder. The decoding part transforms L into $\mathcal{S}' = \mathrm{D}(\mathrm{L})$, where $\mathcal{S}' \in (H/16 \times W/16, C)$. We also use a single class attention layer to aggregate S' into a compact representation f_r and use cross-entropy on f_r as well as L2 loss between S' and S to train the autoencoder. We also use the attention scores from the autoencoder network as key point detection scores to extract top local descriptors, which we found are much cleaner than those from the ViTGaL model.

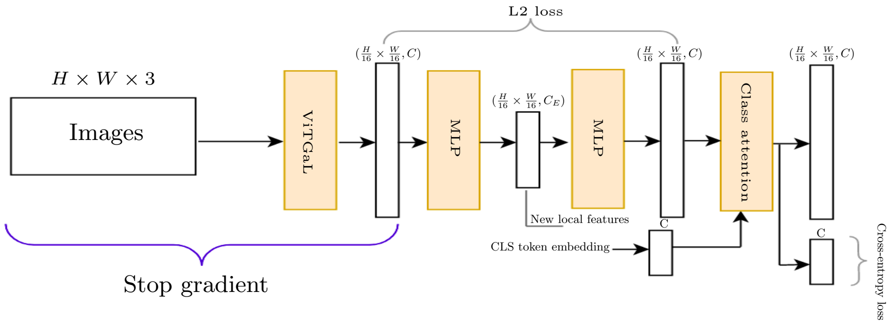

L2 loss
$H \times W \times 3$
Images
ViTGaL
$(\frac{H}{16} \times \frac{W}{16}, C)$
MLP
$(\frac{H}{16} \times \frac{W}{16}, C_E)$
New local features
MLP
$(\frac{H}{16} \times \frac{W}{16}, C)$
C
CLS token embedding
Class attention
$(\frac{H}{16} \times \frac{W}{16}, C)$
C
Cross-entropy loss
Stop gradient

Fig. 3: The architecture of our dimensionality reduction model. It is an autoencoder model, where both encoder and decoder are a simple MLP. L2 loss and cross-entropy loss are used to train the network. The trained ViTGaL to extract high-dimensionality local features for reduction is fixed during training.

Each extracted local feature for the image patch at position h, w is represented with a token embedding $l_{h,w} \in \mathcal{L}$ and its corresponding keypoint detection score $a_{h,w}$ obtained by averaging the attention score from multiple heads at position h,w. These token embeddings are treated as local descriptors for the reranking stage, where their location in the input image is set to the center of their corresponding patch.

3.2 Training Objective

ViTGaL. Following state-of-the-art methods in [9,55], we propose to train our model using only image-level labels. To train both global and local features for the ViTGaL model, we use the ArcFace margin loss [12], where we add only one L2-normalized N class prediction head $W \in R^{C \times N}$

$$L = -log(\frac{exp(\gamma \times AF(\omega_t^T f_g, 1)}{\sum_i exp(\gamma \times AF(\omega_i^T f_g, y_i)))}) \tag{1}$$

where w_i is the ith row of W and f_g is the L2- normalized of the global feature output from the ViTGaL model, y is the one-hot label vector, and t is the index of the ground-truth class($y_t = 1$). γ is a learnable scalar, and AF denotes the ArcFace-adjusted cosine similarity, calculated as follows:

$$AF(s, c) = \begin{cases} cos(acos(s) + m), & \text{if c = 1} \\ s, & \text{if c = 0} \end{cases} \tag{2}$$

where s is the cosine similarity, m is the ArcFace margin, and c is a binary value which c=1 means this is a ground-truth class.

Autoencoder model. We follow [9] and use two losses: the mean-squared error regression loss and cross-entropy loss to train our autoencoder model. First, the mean-squared error regression loss measures how well the autoencoder reconstructs S:

$$L_r(\mathcal{S}, \mathcal{S}') = \frac{1}{H/16 \times W/16 \times C} \sum_{h,w} (\|\mathcal{S}_{h,w} - \mathcal{S}'_{h,w}\|)^2 \tag{3}$$

A cross-entropy loss is also used on top of the aggregation vector f_r of S' using a single class attention layer:

$$L_c(f_r, k) = -log(\frac{exp(v_t^T f_r + b_t)}{\sum_i exp(v_i^T f_r + b_i)}) \tag{4}$$

where v_i, b_i is the classifier weight and bias for class i and t is the ground-truth class for f_r. The total loss for autoencoder model is given by : $L_a = L_c + \lambda L_r$, , where λ is a loss weight

4 Experiments

4.1 Experimental Setup

Training dataset. We use the cleaned version of Google landmarks dataset V2 (GLDv2-cleaned) [51] for training. It contains a total of 1,580,470 images and 81,313 classes. It is a subset of the bigger but more noisy dataset Google land-marks dataset V2 (the original dataset contains 5M images of 200K different landmarks). Google developed the original dataset to raise the challenges faced by the landmark identification system under real industrial scenarios. The cleaned version is built by the competitors from Google Landmark Retrieval Competition 2019, as they found the original dataset is too noisy.

Evaluation datasets and metrics. We use $\mathcal{R}$Oxf, $\mathcal{R}$Par, $\mathcal{R}$Oxf+$\mathcal{R}$1M, $\mathcal{R}$Par+$\mathcal{R}$1M to evaluate our method. $\mathcal{R}$Oxf, $\mathcal{R}$Par [37] are a special version of the original Oxford5k [35] and Paris6k [36] datasets with revisited annotations. Both datasets contain 70 query images and additionally include 4993 and 6322 database images, respectively. $\mathcal{R}$1M [37]refers to the dataset with additional 1M distractor images for evaluating large-scale retrieval. Mean Average Precision (mAP) is used to evaluate the performance of our method on the Medium and Hard splits of all datasets.

Implementation details. We trained our model using the GLDv2-cleaned dataset. For a fair comparison with the state-of-the-art methods in [55,9], we follow them and randomly divide 80% of the dataset for training and the rest 20% for validation. We use XCiT-S12/16 and XCiT-S24/16 as our XCIT backbone models since they have a compatible number of parameters to the Restnet50 and Resnet101 backbones used in state-of-the-art CNN-based methods [55,9]. Our models are initialized from ImageNet pre-trained weights. The image first undergoes augmentation by randomly rotating, shifting, scaling, and cropping, then resizing to 512 × 512 resolution. We use a batch size of 64 and train our model using 2 Tesla V100 GPUs with 32GB memory per card for 40 epochs. Adam optimizer and cosine learning rate decay strategy are adopted for training. We train our model with two warming-up epochs with the initial learning rate of $3e^{-5}$. We then train our model for additional 38 epochs, and the maximum learning rate is set as $5e^{-5}$. We set the ArcFace margin m = 0.1, the ArcFace scale $\gamma = 30$ and the loss weight for L_r to $\lambda = 10$.

As for feature extraction, while previous works [9,55] used multiple scales to extract global features, we use only a single scale. For local features extraction, since using a single scale image representation only can't produce enough local features for reranking, we used five scales, i.e., 0.3535, 0.5, 0.7071, 1.0, 1.4142 for the extraction, although the experiment using only single scale still gives a good result. Local features are selected based on their attention scores. We choose a maximum of 1k local features with the highest attention score. A minimum attention score threshold τ is also used, where we set τ to the median attention score in the last iteration of training following [9]. For local features matching, we use RANSAC [18] with an affine model. We follow [9] and tune the RANSAC parameters on $\mathcal{R}$Oxf, $\mathcal{R}$Par, then the best parameters are fixed for experiments on $\mathcal{R}$Oxf+$\mathcal{R}$1M, $\mathcal{R}$Par+$\mathcal{R}$1M. The top 100 ranked images from the first stage are considered for reranking, where the reranking is based on the number of inliers.

4.2 Comparison with the State-of-the-Art

mAP comparision We compare our model with the state-of-the-art methods in Table 1. All methods are tested on $\mathcal{R}$oxf and $\mathcal{R}$par datasets (and their large-scale versions $\mathcal{R}$oxf+1M, $\mathcal{R}$par+1M), with both Medium and Hard evaluation protocols. We follow previous works [9,55] and divide the previous state-of-the-art methods into three groups: (A) local feature aggregation and re-ranking; (B) global feature similarity search; (C) global feature search followed by re-ranking with local feature matching and spatial verification (SP). Note that although DOLG [55] proposed fusing global and local features into compact image representations, the search is conducted on the fused global features, so we grouped DOLG into the group (B).

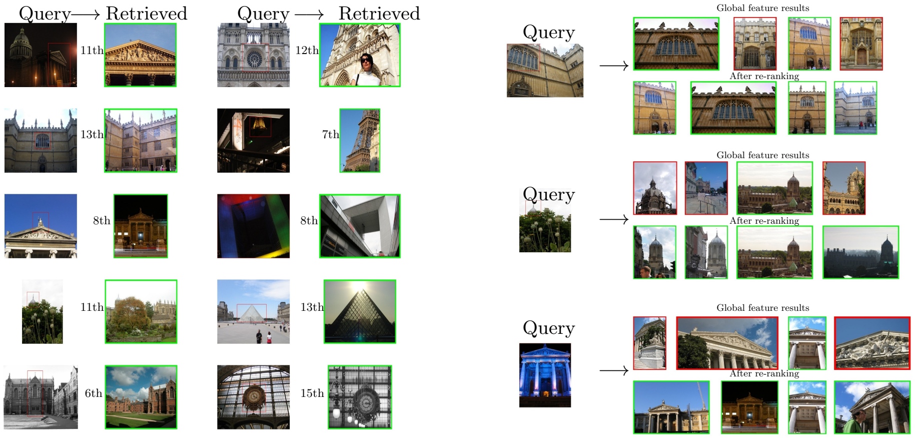
Compared to methods in group B, our ViTGaL global feature variants are significantly better in all cases. In a standard evaluation setting, where query images are cropped [37], our model strongly outperforms state-of-the-art DELG trained on the GLDv2-clean dataset. For example, with the XCiT-S12/16 backbone (roughly the same parameters of ResNet50), the mAP is 79.64% v.s. 73.60% on $\mathcal{R}$oxf-Medium and 62.03% v.s. 51.00% on $\mathcal{R}$oxf-Hard. The gap is more significant in large-scale setting, with 13.48% absolute improvement on $\mathcal{R}$oxf-Hard+1M and 8.32% on $\mathcal{R}$oxf-Medium+1M. Our model with a global feature only also outperforms DELG with the second reranking stage. In evaluation settings where query images are not cropped, our ViTGaL global feature variants also outperform the state-of-the-art one-stage retrieval model DOLG. Note that unlike other methods, which use three [9] or five [55] image scales, our model use only a single image scale to perform retrieval.

Table 1: mAP comparison against the state-of-the-art retrieval methods on the $\mathcal{R}$oxf and $\mathcal{R}$par datasets (and their large-scale versions $\mathcal{R}$oxf+1M/$\mathcal{R}$par+1M), with both Medium and Hard evaluation protocols. ⋆ means feature quantization is used, and "†" means second-order loss is added. "GLDv1" and "GLDv2-clean" mark the difference in the training dataset. ˆdenotes evaluations where queries are not cropped. State-of-the-art performances and ours are marked bold.

Method	Medium				Hard			
	$\mathcal{R}$oxf	+1M	$\mathcal{R}$par	+1M	$\mathcal{R}$oxf	+1M	$\mathcal{R}$par	+1M
(A) Local feature aggregation + re-ranking								
HesAff-rSIFT-ASMK⋆ +SP[46]	60.60	46.80	61.40	42.30	36.70	26.90	35.00	16.80
HesAff-HardNet-ASMK⋆ +SP[31]	65.60	-	65.20	-	41.40	-	38.50	-
DELF-ASMK⋆ +SP[34,37]	67.80	53.80	76.90	57.30	43.10	31.20	55.40	26.40
DELF-R-ASMK⋆ +SP[45]	76.00	64.00	80.20	59.70	52.40	38.10	58.60	29.40
R50-How-ASMK,n=2000[47]	79.40	65.80	81.60	61.80	56.90	38.90	62.40	33.70
R50-MDA-ASMK[51]⋆	**81.80**	**68.70**	**83.30**	**64.70**	**62.20**	**45.30**	**66.20**	**38.90**
(B) Global features								
R101-R-MAC[19]	60.90	39.30	78.90	54.80	32.40	12.50	59.40	28.00
R101-GeM ↑ [44]	65.30	46.10	77.30	52.60	39.60	22.20	56.60	24.80
R101-GeM-AP[41]	67.50	47.50	80.10	52.50	42.80	23.20	60.50	25.10
R101-GeM-AP (GLDv1)[41]	66.30	-	80.20	-	42.50	-	60.80	-
R152-GeM[38]	68.70	-	79.70	-	44.20	-	60.30	-
ResNet101-GeM+SOLAR†[32]	69.90	53.50	81.60	59.20	47.90	29.90	64.50	33.40
R50-DELG[9]	69.70	55.00	81.60	59.70	45.10	27.80	63.40	34.10
R50-DELG (GLDv2-clean)[9]	**73.60**	**60.60**	**85.70**	**68.60**	**51.00**	**32.70**	**71.50**	**44.40**
R101-DELG[9]	73.20	54.80	82.40	61.80	51.20	30.30	64.70	35.50
R101-DELG(GLDv2-clean)[9]	**76.30**	**63.70**	**86.60**	**70.60**	**55.60**	**37.50**	**72.40**	**46.90**
XCiT-S12/16-ViTGaL(GLDv2-clean)(Ours)	**79.64**	**68.92**	**91.58**	**80.91**	**62.03**	**46.18**	**81.98**	**61.93**
XCiT-S24/16-ViTGaL(GLDv2-clean)(Ours)	**79.63**	**69.39**	**91.36**	**81.34**	**61.34**	**46.30**	**81.46**	**62.94**
R50-DOLG (GLDv2-clean)ˆ[55]	**80.50**	**76.58**	**89.81**	**80.79**	**58.82**	**52.21**	**77.70**	**62.83**
R101-DOLG (GLDv2-clean)ˆ[55]	**81.50**	**77.43**	**91.02**	**83.29**	**61.10**	**54.81**	**80.30**	**66.69**
XCiT-S12/16-ViTGaL(GLDv2-clean)ˆ(Ours)	**83.55**	**77.13**	**92.12**	**83.14**	**64.94**	**53.64**	**83.38**	**66.42**
XCiT-S24/16-ViTGaL(GLDv2-clean)ˆ(Ours)	**84.42**	**77.95**	**92.53**	**84.01**	**65.89**	**55.37**	**83.60**	**67.76**
(C) Global feature + Local features re-ranking								
R101-GeM↑+DSM[44]	65.30	47.60	77.40	52.80	39.20	23.20	56.20	25.00
R50-DELG[9]	75.10	61.10	82.30	60.50	54.20	36.80	64.90	34.80
R50-DELG(GLDv2-clean)[9]	**78.30**	**67.20**	**85.70**	**69.60**	**57.90**	**43.60**	**71.00**	**45.70**
R101-DELG[9]	78.50	62.70	82.90	62.60	59.30	39.30	65.50	37.00
R101-DELG(GLDv2-clean)[9]	**81.20**	**69.10**	**87.20**	**71.50**	**64.00**	**47.50**	**72.80**	**48.70**
XCiT-S12/16-ViTGaL+Autoencoder(GLDv2-clean)(Ours)	**83.17**	**74.04**	**91.51**	**81.42**	**66.72**	**51.62**	**80.87**	**62.14**
XCiT-S24/16-ViTGaL+Autoencoder(GLDv2-clean)(Ours)	**82.37**	**74.26**	**91.38**	**81.84**	**64.30**	**52.27**	**80.40**	**62.68**
XCiT-S12/16-ViTGaL+Autoencoder(GLDv2-clean)ˆ(Ours)	**85.62**	**78.79**	**92.34**	**83.60**	**69.37**	**58.09**	**83.25**	**66.97**
XCiT-S24/16-ViTGaL+Autoencoder(GLDv2-clean)ˆ(Ours)	**86.66**	**80.27**	**92.66**	**84.58**	**70.57**	**59.56**	**83.20**	**68.16**

For setup (C), we used both global and local features for retrieval. Local feature re-ranking boosts performance substantially for ViTGaL, especially in large-scale settings: gains of up to 6% (in $\mathcal{R}$oxf+Hard+1M). Our retrieval results also outperform the previous state-of-the-art DELG significantly, by more than 15 % on $\mathcal{R}$paris+Hard+1M and 8 % on $\mathcal{R}$oxf+Hard+1M. ViTGaL also outperforms local feature aggregation results from setup (A) in all cases, establishing a new state-of-the-art across the board.

Qualitative results. We showcase the retrieval results of our model in Figures 4a and 4b. Figure 4a illustrates the challenging cases where the gallery images show significant lighting changes and extreme viewpoint differences. These images are still capable of achieving relatively high ranks due to the effectiveness of our global feature, which captures the similarity well even in such challenging scenarios.



(a) Global feature retrieval.

(b) Local feature re-ranking.

Fig. 4: Sample of ViTGaL results on $\mathcal{R}$oxf-Hard+1M and $\mathcal{R}$paris-Hard+1M. **(a)** Examples of challenging, high-ranked relevant retrieved images under the global feature retrieval. **(b)** Examples illustrating accuracy improvements using local features matching for re-ranking. For each query (left), two rows are presented on the right, the top one showing results based on global feature similarity and the bottom one showing results after re-ranking. All rows on the right show the top four retrieval results.

Figure 4b shows the effect of local feature re-ranking of our methods, where global features alone are not enough for high retrieval performance. Global features tend to retrieve images with a generally similar appearance but do not always depict the same object of interest. This can be significantly improved with local feature re-ranking, allowing stricter matching selectivity.

To further illustrate the power of the local feature of ViTGaL, we present qualitative results of local feature matching in Figure 5 and Figure 6. These visualizations depict the final obtained correspondences after RANSAC. We show the matching results between one query and two gallery images per row, where images with lines connecting the corresponding key points are on the figure's right. Figure 5 showcases the robustness of VitGaL in extreme cases such as strong viewpoint and illumination changes, where matches can be obtained across different scales, in occlusion cases, and in day-vs-night scenarios. Figure 6 presents the matching between images of different scenes/objects: matches are still found due to the similarity in patterns between query and index images (e.g., similar windows, arches, or roofs). Nevertheless, these do not affect retrieval results much because the number of inliers is low.

4.3 Ablation Experiments

We conducted experiments using the XCiT-S12/16 backbone to verify some of our design choices empirically.

Verification of the multi-atrous convolution. A multi-atrous convolution block is added to our model to simulate the image pyramid used in the standard retrieval algorithms. We provide experimental results to validate the contribution of the multi-atrous convolutions by removing them from the ViTGaL model. The results is shown in Table 2. It can be seen that adding the multi-atrous convolutions helps to improve the overall performance significantly. The mAP is improved from 57.80% to 62.03% and 79.18% to 81.98% on $\mathcal{R}$oxf-Hard and $\mathcal{R}$par-Hard, respectively. Moreover, our model can achieve state-of-the-art results using only single-scale representation at inference time. Table 3 reports the performance of the global retrieval of our model under different numbers of image scales. The scale rates are 0.7071, 1.0, 1.4142; 0.3535, 0.5, 0.7071, 1.0, 1.4142; 0.25, 0.3535, 0.5, 0.7071, 1.0, 1.4142, 2 for 3-scale, 5-scale and 7-scale setting respectively. To fuse these multi-scale features, we follow the previous works in [9,55] by firstly L2 normalizing them, then averaging the normalized features, and finally applying again an L2 normalization to produce the final descriptor. We can find from the empirical results that using single-scale only performs the best among four multi-scale settings. Such experimental results are also within our expectations. With multi-atrous convolution, we can simulate the image pyramid within the feature space directly. These results validate the effectiveness of the multi-atrous convolution in our model.

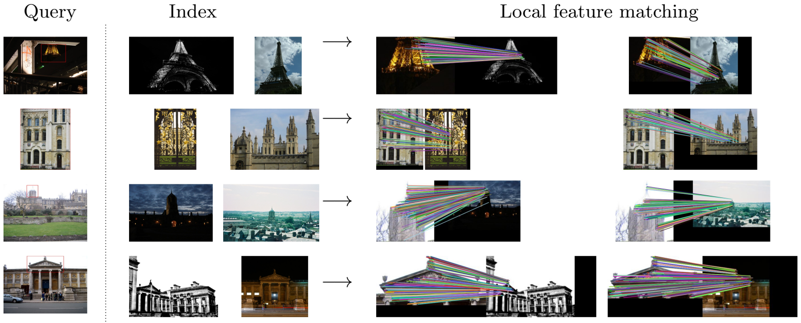


Fig. 5: Example of local feature matches for image pairs depicting the same objects/scenes.

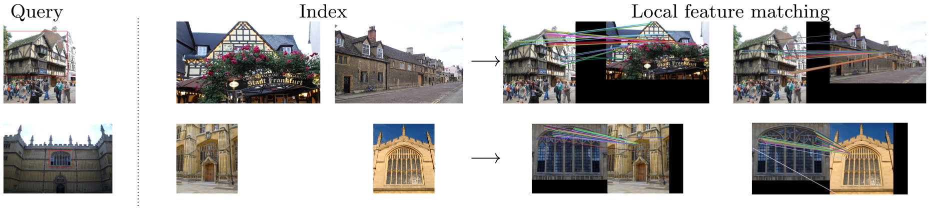


Fig. 6: Example of local feature matches for image pairs depicting different objects/scenes.

Table 2: Ablation experiment on the multi-atrous convolution.

Config	$\mathcal{R}$oxf-M	$\mathcal{R}$oxf-H	$\mathcal{R}$par-M	$\mathcal{R}$par-H
w/o multi-atrous convolution	78.77	57.8	90.4	79.18
Full model	**79.64**	**62.03**	**91.58**	**81.98**

Re-ranking experiment. Table 4 compares local features for re-ranking under different number of scales. The scaling rates are exactly the same as those used in the previous experiments presented in Table 3. All scales receive the identical retrieval short list of 100 images for re-ranking from the XCiT-S12/16-ViTGaL global retrieval result. For a fair comparison, all scales use 1k features and 2k RANSAC iterations. We also tune the matching hyperparameters separately for each scale. The tuning hyperparameters are the distance threshold for selecting correspondences, RANSAC reprojection error threshold, and RANSAC homography confidence. Unlike the global retrieval experiments where using a single-scale only shows superior performance, the best result in re-ranking is observed with the 3-scale, and 5-scale settings since using a single scale only can't produce enough local features (average number of local features in $\mathcal{R}$oxf dataset are 250.6 and 876.7 for 1-scale and 5-scales setting respectively). Nevertheless, using a single scale only gives competitive results compared to the best scale setting, with the largest gap with the best scale among the four evaluation benchmark being only 1.7 % (at $\mathcal{R}$oxf-H).

Latency and memory In Table 5, we list the memory footprint and extraction latency required by different methods for $\mathcal{R}$1M; corresponding to the three settings from Table 1. Similar to DELG, using ViTGaL for joint extraction allows

Table 3: Global retrieval results of ViTGaL under different multi-scale settings.

Number of scales	$\mathcal{R}$oxf-M	$\mathcal{R}$oxf-H	$\mathcal{R}$par-M	$\mathcal{R}$par-H
7-scale	77.6	60.01	91.19	80.74
5-scale	77.73	59.90	91.18	80.75
3-scale	79.28	60.53	**91.68**	**82.18**
1-scale	**79.64**	**62.03**	91.58	81.98

Table 4: Re-ranking results of ViTGaL under different multi-scale settings.

Number of scales	$\mathcal{R}$oxf-M	$\mathcal{R}$oxf-H	$\mathcal{R}$paris-M	$\mathcal{R}$paris-H
7-scale	82.29	65.10	91.34	80.79
5-scale	**83.17**	**66.72**	91.51	80.87
3-scale	83.15	66.27	**91.61**	**81.20**
1-scale	82.53	65.02	91.38	80.73

a significant speedup over using two separate local and global models. Moreover, our model performance is much better than DELG, especially for a single-scale setting, where our model is superior in both storage and speed (3x times faster than DELG and requires only 25 % storage of DELG).

Table 5: Feature extraction latency and database memory requirements for different image retrieval models. Latency is measured on an NVIDIA Tesla P100 GPU.

Method	Extraction latency (ms)	Memory (GB)	
		$\mathcal{R}$Oxf+1M	$\mathcal{R}$Par+1M
(A) Local feature aggregation			
DELF-R-ASMK* [45]	2260	27.6	-
(B) Global features			
R50-GeM [38]	100	7.7	7.7
R101-GeM [38]	175	7.7	7.7
(C) Unified global + local features			
R50-DELG [9]	211	485.5	486.2
R101-DELG [9]	383	485.9	486.6
XCiT-S12/16-ViTGaL+Autoencoder[ours]	158	420.1	420.8
XCiT-S12/16-ViTGaL+Autoencoder (1 scale global & local) [ours]	63	120.1	120.2
XCiT-S24/16-ViTGaL+Autoencoder[ours]	302	420.5	421.2
XCiT-S12/16-ViTGaL+Autoencoder (1 scale global & local) [ours]	87	120.3	120.4

5 Conclusions

In this paper, we make the first attempt to learn a model that enables joint extraction of local and global image features with ViTs, referred to as ViTGaL. The model is based on an XCiT backbone, leveraging multi-atrous convolutions to simulate the spatial feature pyramid used in the standard image retrieval algorithms for high-performing retrieval using a single-scale image representation only. The entire network can be trained end-to-end using image-level labels. We also use an autoencoder to reduce the dimension of local features for an effective re-ranking step. Extensive experiments demonstrate the superior performance of our method on image retrieval, achieving state-of-the-art performance on the Revisited Oxford and Revisited Paris datasets.

References

1. Ali, A., Touvron, H., Caron, M., Bojanowski, P., Douze, M., Joulin, A., Laptev, I., Neverova, N., Synnaeve, G., Verbeek, J., et al.: Xcit: Cross-covariance image transformers. Advances in neural information processing systems **34** (2021)
2. Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN architecture for weakly supervised place recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
3. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6836–6846 (2021)
4. Avrithis, Y., Tolias, G.: Hough pyramid matching: Speeded-up geometry re-ranking for large scale image retrieval. International Journal of Computer Vision (02 2014)
5. Balntas, V., Riba, E., Ponsa, D., Mikolajczyk, K.: Learning local feature descriptors with triplets and shallow convolutional neural networks. In: Bmvc. p. 3 (2016)
6. Barroso-Laguna, A., Riba, E., Ponsa, D., Mikolajczyk, K.: Key. net: Keypoint detection by handcrafted and learned cnn filters. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5836–5844 (2019)
7. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Speeded-up robust features (surf). Computer Vision and Image Understanding **110**(3), 346–359 (2008)
8. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding. arXiv preprint arXiv:2102.05095 **2**(3), 4 (2021)
9. Cao, B., de Araújo, A.F., Sim, J.: Unifying deep local and global features for image search. In: ECCV (2020)
10. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
11. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 1, pp. 539–546. IEEE (2005)
12. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
13. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 337–33712 (2018). https://doi.org/10.1109/CVPRW.2018.00060
14. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 224–236 (2018)
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
16. Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T.: D2-net: A trainable cnn for joint detection and description of local features. arXiv preprint arXiv:1905.03561 (2019)
17. Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., Feichtenhofer, C.: Multiscale vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6824–6835 (2021)

18. 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)
19. Gordo, A., Almazán, J., Revaud, J., Larlus, D.: End-to-end learning of deep visual representations for image retrieval. CoRR **abs/1610.07940** (2016)
20. He, K., Lu, Y., Sclaroff, S.: Local descriptors optimized for average precision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 596–605 (2018)
21. Hinton, G.E.: Connectionist learning procedures. In: Machine learning, pp. 555–610. Elsevier (1990)
22. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact representation. In: IEEE Conf. Comput. Vis. Pattern Recognit. pp. 3304–3311 (2010)
23. Jégou, H., Chum, O.: Negative evidences and co-occurences in image retrieval: The benefit of pca and whitening. pp. 774–787 (10 2012)
24. Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P., Schmid, C.: Aggregating local image descriptors into compact codes. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(9), 1704–1716 (2012)
25. Li, Z., Wang, X., Liu, X., Jiang, J.: Binsformer: Revisiting adaptive bins for monocular depth estimation. arXiv preprint arXiv:2204.00987 (2022)
26. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al.: Swin transformer v2: Scaling up capacity and resolution. arXiv preprint arXiv:2111.09883 (2021)
27. 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: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
28. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**(2), 91–110 (nov 2004)
29. Luo, Z., Shen, T., Zhou, L., Zhang, J., Yao, Y., Li, S., Fang, T., Quan, L.: Contextdesc: Local descriptor augmentation with cross-modality context. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2527–2536 (2019)
30. Mishchuk, A., Mishkin, D., Radenović, F., Matas, J.: Working hard to know your neighbor's margins: Local descriptor learning loss. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 4829–4840. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
31. Mishkin, D., Radenović, F., Matas, J.: Repeatability is not enough: Learning affine regions via discriminability. In: ECCV (2018)
32. Ng, T., Balntas, V., Tian, Y., Mikolajczyk, K.: Solar: second-order loss and attention for image retrieval. In: European Conference on Computer Vision. pp. 253–270. Springer (2020)
33. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 2, pp. 2161–2168 (2006). https://doi.org/10.1109/CVPR.2006.264
34. Noh, H., Araujo, A., Sim, J., Weyand, T., Han, B.: Large-scale image retrieval with attentive deep local features. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 3476–3485 (2017). https://doi.org/10.1109/ICCV.2017.374
35. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: 2007 IEEE conference on computer vision and pattern recognition. pp. 1–8. IEEE (2007)

36. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In: 2008 IEEE conference on computer vision and pattern recognition. pp. 1–8. IEEE (2008)
37. Radenović, F., Iscen, A., Tolias, G., Avrithis, Y., Chum, O.: Revisiting oxford and paris: Large-scale image retrieval benchmarking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5706–5715 (2018)
38. Radenovic, F., Tolias, G., Chum, O.: Fine-tuning CNN image retrieval with no human annotation. IEEE Trans. Pattern Anal. Mach. Intell. **41**(7), 1655–1668 (2019). https://doi.org/10.1109/TPAMI.2018.2846566, https://doi.org/10.1109/TPAMI.2018.2846566
39. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? Advances in Neural Information Processing Systems **34** (2021)
40. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12179–12188 (2021)
41. Revaud, J., Almazán, J., Rezende, R.S., Souza, C.R.d.: Learning with average precision: Training image retrieval with a listwise loss. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5107–5116 (2019)
42. Revaud, J., Weinzaepfel, P., De Souza, C., Pion, N., Csurka, G., Cabon, Y., Humenberger, M.: R2d2: repeatable and reliable detector and descriptor. arXiv preprint arXiv:1906.06195 (2019)
43. Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From coarse to fine: Robust hierarchical localization at large scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12716–12725 (2019)
44. Simeoni, O., Avrithis, Y., Chum, O.: Local features and visual words emerge in activations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
45. Teichmann, M., Araujo, A., Zhu, M., Sim, J.: Detect-to-retrieve: Efficient regional aggregation for image search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5109–5118 (2019)
46. Tolias, G., Avrithis, Y., Jégou, H.: Image search with selective match kernels: aggregation across single and multiple images. International Journal of Computer Vision **116**(3), 247–261 (2016)
47. Tolias, G., Jenicek, T., Chum, O.: Learning and aggregating deep local descriptors for instance-level recognition. In: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I. p. 460–477. Springer-Verlag, Berlin, Heidelberg (2020)
48. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of cnn activations (2016)
49. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5265–5274 (2018)
50. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 568–578 (2021)
51. Wu, H., Wang, M., Zhou, W., Li, H.: Learning deep local features with multiple dynamic attentions for large-scale image retrieval. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11396–11405 (2021). https://doi.org/10.1109/ICCV48922.2021.01122

52. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34** (2021)
53. Yan, H., Zhang, C., Wu, M.: Lawin transformer: Improving semantic segmentation transformer with multi-scale representations via large window attention. arXiv preprint arXiv:2201.01615 (2022)
54. Yandex, A.B., Lempitsky, V.: Aggregating local deep features for image retrieval. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 1269–1277 (2015)
55. 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: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11772–11781 (October 2021)
56. Yi, K.M., Trulls, E., Lepetit, V., Fua, P.: Lift: Learned invariant feature transform. In: European conference on computer vision. pp. 467–483. Springer (2016)
57. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605 (2022)
58. Zhang, Y., Jin, R., Zhou, Z.H.: Understanding bag-of-words model: a statistical framework. International Journal of Machine Learning and Cybernetics **1**(1), 43–52 (2010)