UltraGlobal: An Enhanced Approach to Image Retrieval Using Global Features

Xuanlang Dai[1], Pengfei Huang[2], Shicheng Wang[3], Zhiqi Zhang[4,*] and Mingyang Gao[5]
{1342465252@xjtu.edu.cn[1], u202341843@xs.ustb.edu.cn[2], 1878094894@qq.com[3],
zhangzq2023@mails.jlu.edu.cn[4], gaomingyang@bit.edu.cn[5]}

Xi'an Jiaotong University, Xi'an, 710049, China[1]
University of Science and Technology Beijing, Beijing, 100083, China[2]
Xi'an Jiaotong University, Xi'an, 710049, China[3]
Jilin University, Jilin, 130015, China[4]
Beijing Institute of Technology, Beijing, 100081, China[5]
*corresponding author

Abstract. Image retrieval is an important task in vision, and for a long time, the research has focused on traditional algorithms or DL methods, but neglected the better measurement of feature extraction brought by the combination of the two, based on this, we propose UltraGlobal, an Image retrieval method that uses DL method to extract features and encoding traditional algorithms, and our main contributions are: 1) the introduction of PANet in the feature extraction stage; 2) adopt long Global descriptors and improve GeM pooling; 3) NetVLAD(VLAD) was introduced as an encoding layer; Experimental results demonstrate that UltraGlobal significantly outperforms existing methods on standard benchmarks, showcasing exceptional scalability and precision. This approach offers a more efficient and accurate solution for image retrieval systems. Code: https://github.com/Lennox-Dai/UltraGlobal.

Keywords: image retrieval, visual search, SuperGlobal, global feature, deep learning

1 Introduction

Image retrieval systems are crucial for various applications, such as digital asset management and visual search engines. These systems aim to identify and retrieve images from extensive databases that are similar to a given query image. The retrieval process is typically divided into two key stages. Initially, a fast and efficient method sorts the database images based on their estimated high-level similarity to the query. This stage is essential for narrowing down the vast pool of potential matches to a manageable subset. Subsequently, in the reranking stage, the top candidates

CONF-MLA 2024, November 21, Adana, Turkey

DOI 10.4108/eai.21-11-2024.2354612

undergo a more detailed and computationally intensive matching process against the query image, refining the initial results to produce a more accurate ranked list [1, 2, 3, 4].

In contemporary implementations, the initial stage frequently employs deep learning-based global features. These methods have gained significant traction in recent years due to their robustness and efficiency [5, 6, 7, 8]. On the other hand, the reranking stage often relies on geometric matching of local image features [1, 2, 9, 10]. This technique provides valuable information about the spatial consistency between the query and the database images, enhancing the accuracy of the retrieval results.

Recent trends in image retrieval research have focused on utilizing advanced matching processes during the reranking stage. Techniques such as transformers [11] and 4D correlation networks [12] have demonstrated remarkable improvements in retrieval quality. However, these sophisticated methods come with significant drawbacks, including increased reranking latency (several seconds per query) and substantial memory requirements (over 1MB per database image). These limitations pose significant challenges when scaling to large repositories.

Our research directly addresses these limitations by introducing a novel method that relies entirely on global image features for both retrieval stages. Additionally, we revisit pooling techniques, proposing new modules to enhance global feature extraction. Our method, **UltraGlobal**, is illustrated in fig. 1 and brings the following innovations to the field of image retrieval:

- **Enhanced Feature Extraction with PANet.** We have incorporated PANet to extract comprehensive global features, significantly improving feature representation and capturing more detailed information about the images.
- **Multiple Improvements to GeMP Modules.** We propose several enhancements to the GeMP (Generalized Mean Pooling) module, enabling the extraction of multiple, richer features. These improvements allow the system to capture a wider variety of image characteristics, leading to better retrieval performance.
- **Advanced Encoding Techniques with VLAD.** We further refine the feature extraction process by integrating advanced encoding methods such as VLAD. This step improves the system's ability to recognize subtle differences in images, thereby significantly enhancing overall retrieval accuracy [1, 5].

UltraGlobal represents a groundbreaking approach to image retrieval that relies solely on global features throughout the process. This eliminates the need for costly reranking based on local features, significantly improving scalability and efficiency. Experimental results on standard benchmarks highlight the effectiveness of the proposed method, setting new state-of-the-art performance levels and demonstrating substantial improvements over previous approaches [13].

In summary, the unique contribution of UltraGlobal lies in its innovative use of global features, advanced pooling techniques, and sophisticated encoding methods. These advancements collectively lead to a more efficient, scalable, and accurate image retrieval system, addressing key limitations of current state-of-the-art methods.

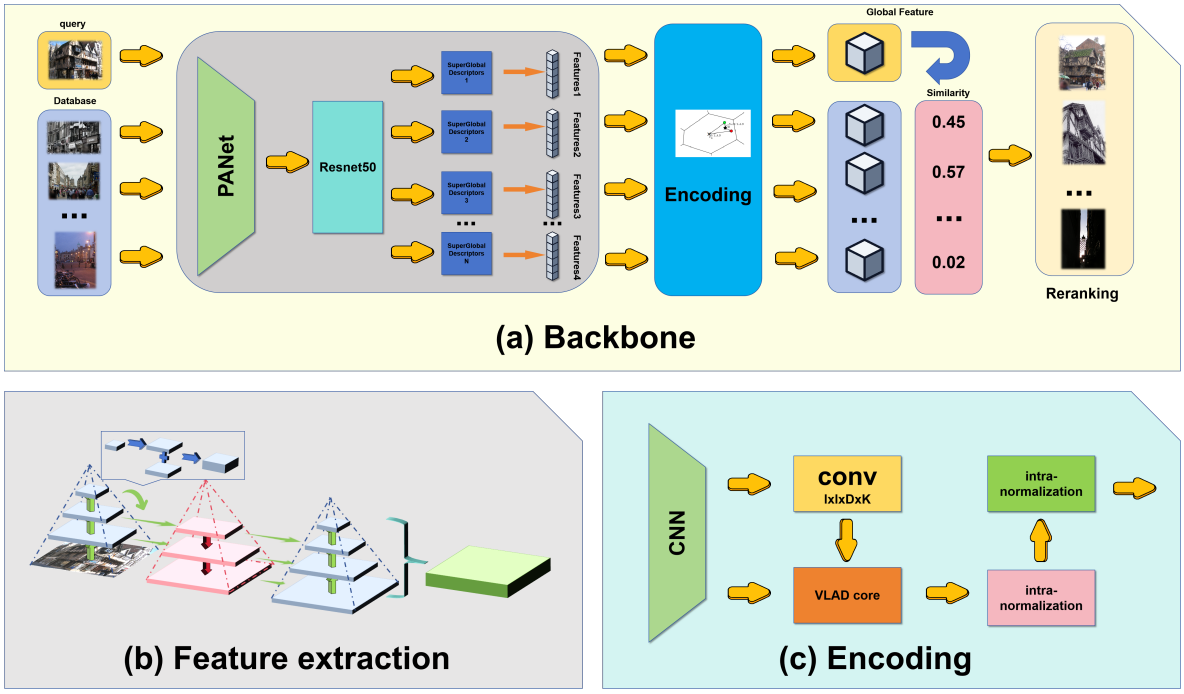

Fig. 1. Overview of UltraGlobal

2 Related Work

2.1 Feature Extraction

SIFT (Scale-Invariant Feature Transform) [14] is a classic feature extraction algorithm identifying key points invariant to scale and rotation, robust against illumination and noise changes. It is widely used in tasks like object recognition, image stitching, and 3D modeling. CVNet [15] is a deep learning-based convolutional neural network that extracts high-level image features, outperforming traditional methods like SIFT in recognition and retrieval tasks. It leverages large datasets for robust, generalizable feature representations. Tokens [16] represent images as sequences of tokens, akin to words in NLP, allowing the use of transformers. This approach captures long-range dependencies and contextual information, improving image retrieval accuracy. Superglobal [17] combines local feature extraction like SIFT with the global representation capabilities of CNNs. This hybrid approach captures fine details and overall context, enhancing robustness and performance in varied scenarios.

2.2 Encoding

Encoding Following the extraction of image features, the incorporation of an encoding step can significantly enhance the efficiency and accuracy of image retrieval systems. Currently, prevalent encoding methods include:

- **Bag of Visual Words (BoVW).** a classical feature aggregation technique that quantifies local features in images and maps them to a visual vocabulary, thereby creating a histogram descriptor.[18, 19]
- **Fisher Vector (FV).** represents a robust feature encoding approach that combines Gaussian Mixture Models (GMM) with Fisher Information Matrices. It generates global descriptors by statistically analyzing the distribution of local features.[20, 21]
- **VLAD (Vector of Locally Aggregated Descriptors).** a feature encoding method that aggregates local features into a compact global image descriptor. It clusters feature descriptors, calculates the residuals relative to the cluster centers, and accumulates these residual vectors to produce the final descriptor. VLAD exhibits exceptional performance in image retrieval tasks, efficiently representing image features.[22]
- **NetVLAD.** a neural network-based feature aggregation method, leverages the advantages of VLAD and deep learning. It produces high-quality global descriptors through end-to-end training.[5]
- **R-MAC (Regional Maximum Activations of Convolutions).** extracts features from multiple regions of an image and applies maximum pooling to generate a global descriptor. This method effectively captures details and spatial information within images.[23]
- **Cross-dimensional Weighting (CroW).** enhances the discriminative power of feature representations by weighting each dimension of the feature map. It is particularly suited for features extracted by deep convolutional neural networks (CNNs).[24]
- **DEFL (Deep Local Feature Learning).** combines local and global features through end-to-end learning to produce robust image descriptors.[25]

2.3 Reranking

Reranking is the process of refining the initial search results in image retrieval by applying a more detailed matching process to the top-ranked images from the initial retrieval. This typically improves the accuracy and relevance of the final ranked list.

Geometric Verification (GV) [19, 26] was once regarded as the most effective reranking method, leveraging the geometric relationships within images for precise matching. However, with the rapid advancements in deep learning technologies, researchers have increasingly adopted more complex and parameter-rich models for reranking. These models include transformers [27] and 4D convolutional neural networks (4D CNNs) [28, 29] with deeply stacked 4D convolution layers. These

advanced models have demonstrated exceptional performance in handling complex visual tasks, significantly improving accuracy and making the reranking process more reliable and precise.

Another important research direction involves the use of lightweight convolutional neural network (CNN) models to extract global features, replacing heavier models. This approach aims to enhance the model's inference capability and response speed while maintaining high accuracy. For example, SuperGlobal [17] is a lightweight model that, through architectural and algorithmic optimization, significantly improves inference efficiency while ensuring high precision. This method offers notable advantages in resource-constrained environments and excels in real-time applications, providing faster response and processing for visual tasks.

3 Proposed Methods

3.1 Enhanced Feature Extraction with PANet

Through data attribution analysis, we discovered that despite methods like SuperGlobal using global feature descriptors to extract global image information, the attention weights are still concentrated on certain local regions. Additionally, we found that during the inference phase, the model does not require retraining, meaning that the feature tensors for each image in the database remain stable and unchanged. Consequently, the impact of model complexity on speed can be considered negligible in practical use, as it only affects the feature extraction speed of query images.

Therefore, we propose to increase model complexity by incorporating PANet between the input image and ResNet to enhance the global receptive field, at the cost of sacrificing some feature extraction speed. The specific model architecture is illustrated in fig. 2.

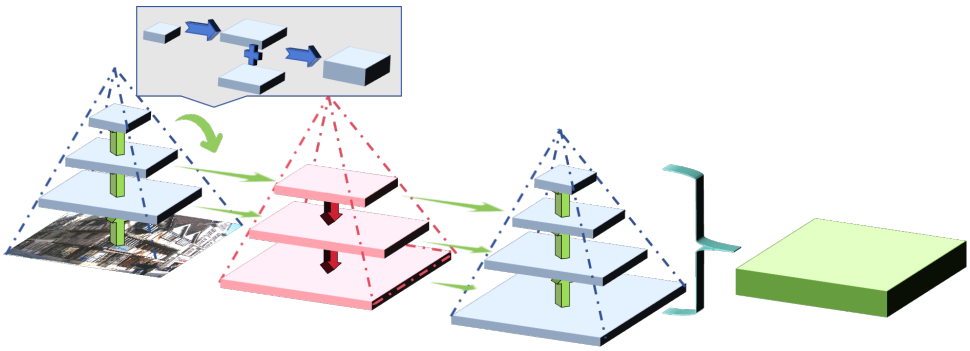

Fig. 2. Enhanced Feature Extraction with PANet

3.2 Multiple Improvements to GeMP Modules

To enhance the feature representation capability of the model, we introduce the Non-Local Block module. This module captures long-range dependencies, helping to enhance global feature

representation and thus improve the overall performance of the model. The method was proposed in [30]. The core idea is to enhance feature representation by computing non-local responses between input features. We build a module based on this idea to improve the performance of our model. Here is the structure and principle:

Feature Transformation. The input feature X undergoes three different convolutional transformations to obtain feature maps G, Θ, and Φ:

$$G = W_g X, \tag{1}$$

$$\Theta = W_\theta X, \tag{2}$$

$$\Phi = W_\phi X \tag{3}$$

Convolutional layers are used here to learn spatial hierarchies of features from the input images, capturing local patterns effectively. Batch normalization layers are applied after these transformations to stabilize and accelerate the training process by normalizing the inputs of each layer, reducing internal covariate shift.

Similarity Calculation. Calculate the similarity matrix f between the feature maps Θ and Φ:

$$f_{ij} = \theta(X_i)^\top \phi(X_j) \tag{4}$$

This matrix makes sense because it captures the pairwise similarities between features, essential for understanding the relationships and dependencies within the data. We ensure that the values are scaled between 0 and 1, representing probabilities that sum to 1, which aids in interpreting the strength of the relationships. We do this by normalizing this similarity matrix using softmax:

$$f_{\text{div\_C}}(i,j) = \frac{\exp(f_{ij})}{\sum_j \exp(f_{ij})} \tag{5}$$

Feature Enhancement. Multiply the normalized similarity matrix $f_{\text{div\_C}}$ with the feature map G to obtain the enhanced feature Y:

$$Y = f_{\text{div\_C}} G \tag{6}$$

This step refines the features by emphasizing the important relationships captured in the similarity matrix.

Feature Reconstruction. Combine the enhanced feature Y with the original input X through a convolutional layer and batch normalization layer to obtain the final output Z:

$$Z = \text{BN}(W_z Y + X) \tag{7}$$

The convolutional layer integrates the enhanced features with the original ones, while the batch normalization layer ensures that the output is normalized, facilitating stable and efficient training.

3.3 Integration with GeM+ Module

To enhance the feature aggregation effect, we integrate the Non-Local Block into the GeM+ module. The workflow of the integrated GeM+ module is detailed as follows: Initially, the input feature undergoes enhancement via the Non-Local Block, which refines the feature representation. Subsequently, the enhanced features are clamped and exponentiated to facilitate effective aggregation. This is followed by global average pooling to aggregate the features. The aggregated feature is produced as the output of the module.

Compared to the original method without Non-Local Block, we introduce Non-Local Block to provide rich feature transformations through multiple convolutional and non-linear operations. Non-Local Block addresses this by computing non-local response. It is able to capture long-range dependencies and enhance the global context understanding, thus retaining more details and improving feature aggregation.

3.4 Advanced Encoding Techniques with VLAD

In this study, we improve the feature representation of the SuperGlobal model by incorporating VLAD (Vector of Locally Aggregated Descriptors) encoding. This approach leverages the advantages of both global and local features, aiming to boost the accuracy and robustness of the image retrieval system. It serves as an efficient feature aggregation technique, which converts local descriptors into a fixed-length global feature vector. The process of VLAD encoding involves the following steps:

- **Cluster Center Initialization.** First, a set of cluster centers $\{c_1, c_2, \ldots, c_K\}$ is predefined from the training data, where c_k represents the k-th cluster center, and K is the number of cluster centers.
- **Computing Soft Assignment.** For each input feature vector x_i, the similarity to all cluster centers is computed. This similarity is then transformed into soft assignment weights α_{ik} using the softmax function, representing the probability that the feature vector x_i belongs to cluster center c_k. The formula for calculating the soft assignment weights is:

$$\alpha_{ik} = \frac{\exp(-\|x_i - c_k\|^2)}{\sum_{j=1}^{K} \exp(-\|x_i - c_j\|^2)} \tag{8}$$

  where $\|x_i - c_k\|$ represents the Euclidean distance between the feature vector x_i and the cluster center c_k.

- **Aggregating Features.** All weighted residuals are summed to obtain the aggregated feature vector v_k for each cluster center:

$$v_k = \sum_i r_{ik} = \sum_i \alpha_{ik}(x_i - c_k) \tag{9}$$

  Then, all aggregated feature vectors v_k are concatenated to form the fixed-length VLAD feature vector:

$$v = [v_1, v_2, \ldots, v_K] \tag{10}$$

- **Normalization.** To improve the stability and comparability of the feature vector, the aggregated feature vector is L2-normalized:

$$\hat{v} = \frac{v}{\|v\|} \tag{11}$$

We believe that integrating VLAD encoding enhances feature representation by considering residuals between local features and cluster centers, improving discriminative power. This combination of global and local information significantly boosts retrieval accuracy, especially for fine-grained distinctions, while maintaining scalability and efficiency suitable for large-scale tasks.

4 Experiments

4.1 Experimental Setup

We conducted an evaluation of our model on the ROxford5k and RParis6k datasets [2]. During the fine-tuning stage, we optimized a combination of ResNet50 and PANet on the GLDv2 dataset [31], where the fine-tuning task was formulated as a standard multi-class classification problem. In the inference stage, we performed image retrieval on query images, which involved searching for images in an existing database that exhibit the highest feature similarity to the query image.

To assess the effectiveness of the proposed method, we selected several common baseline models for comparison, including DELG [8], DOLG [32], CVNet [15], and SuperGlobal [17], which use only global feature retrieval, as well as CVNet and SuperGlobal, which use both global feature retrieval and reranking. We compared models with different network depths to provide a comprehensive evaluation.

For a thorough assessment of our model's performance in image retrieval, we employed the average precision(AP) metric to evaluate the improvements in retrieval effectiveness. Recognizing that using only average precision may not provide a complete measure of model performance, we further visualized the median of the accuracy of the top-ranked images of the queries, offering additional insights into the model's robustness and ensuring the comprehensiveness of the evaluation. Our experimental results are presented below.

To validate the robustness and stability of our model, we extracted the feature extraction component and trained it alongside the original ResNet on a toxic dataset, where "toxic" refers to data that is deliberately designed to significantly degrade the model's performance. The results show that the feature extraction component with the added PANet module better preserved the model's capability compared to the original network.

4.2 Results

4.2.1 Performance Inprovement

Comparing the AP values of our method and some previous methods on ROxford & RParis datasets with different difficulty levels as section 4.2.1, we can see that the effect is improved on datasets with different difficulty levels.

Applying reranking methods demonstrates the best performance improvement compared to methods without the reranking stage, particularly on the ROxford5k-Hard dataset, where the RN50-SuperGlobal method achieved at most over a 10% increase in AP. Furthermore, our approach consistently shows the highest average improvement across different datasets, which underscores the effectiveness of the gempNonLocalBlock module in enhancing performance for this task.

Table 1: Performance comparison of various methods on different datasets and difficulty levels. All reranking methods rerank top 400.

Method	**Medium**		**Hard**	
	ROxf	RPar	ROxf	RPar
Global feature retrieval				
RN50-DELG [8]	73.6	85.7	51.0	71.5
RN101-DELG [8]	76.3	86.6	55.6	72.4
RN50-DOLG [32]	80.5	89.8	58.8	77.7
RN101-DOLG [32]	81.5	91.0	61.1	80.3
RN50-CVNet [15]	81.0	88.8	62.1	76.5
RN101-CVNet [15]	80.2	90.3	63.1	79.1
RN50-SuperGlobal [17]	83.9	90.5	67.7	80.3
RN101-SuperGlobal [17]	85.3	92.1	72.1	83.5
Global feature retrieval + Local feature reranking				
RN50-CVNet	87.9	90.5	75.6	80.2
RN101-CVNet	87.2	91.2	75.9	81.1
SuperGlobal feature retrieval and reranking				
RN50-SuperGlobal	88.8	92.0	77.1	84.4
RN101-SuperGlobal	90.9	93.3	80.2	86.7
RN101-SuperGlobal-gempNonLocalBlock [ours]	**91.4**	**93.4**	**81.0**	**87.0**

In order to understand the effect and robustness of different methods, we made a distribution chart of queries with different AP values fig. 3. The horizontal axis represents the interval of the AP value of the query, and the vertical axis represents the number of queries with AP value in a certain interval. The lower the left side of the line, the higher the right side, the better the effect. The larger peak value of the broken line indicates the more concentrated AP value of the query and the better robustness. It can be seen that the effect and robustness of our method are good, and there are some improvements over the previous method. It can also be seen that reranking significantly enhances the effect and robustness of the method. There is an obvious multi-peak phenomenon in the polyline before reranking, that is, the extremal of the queried quantity appears in multiple AP value intervals, the polyline is scattered, and the upward trend is not as obvious as that after reranking.

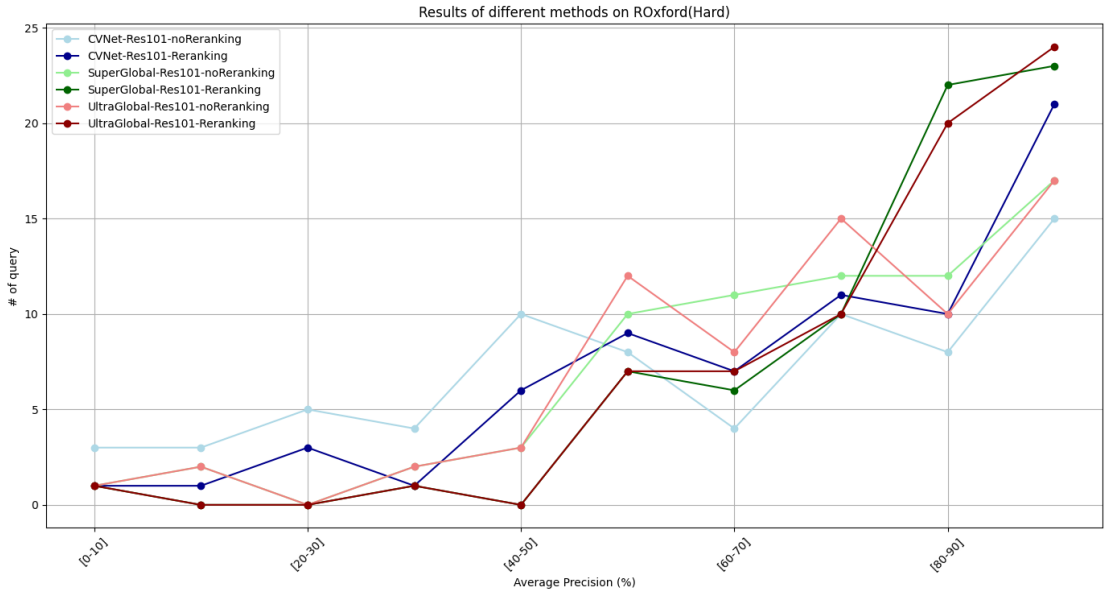

Fig. 3. Distribution of AP values for queries. Using ROxford(Hard) dataset. The number of queries is 70.

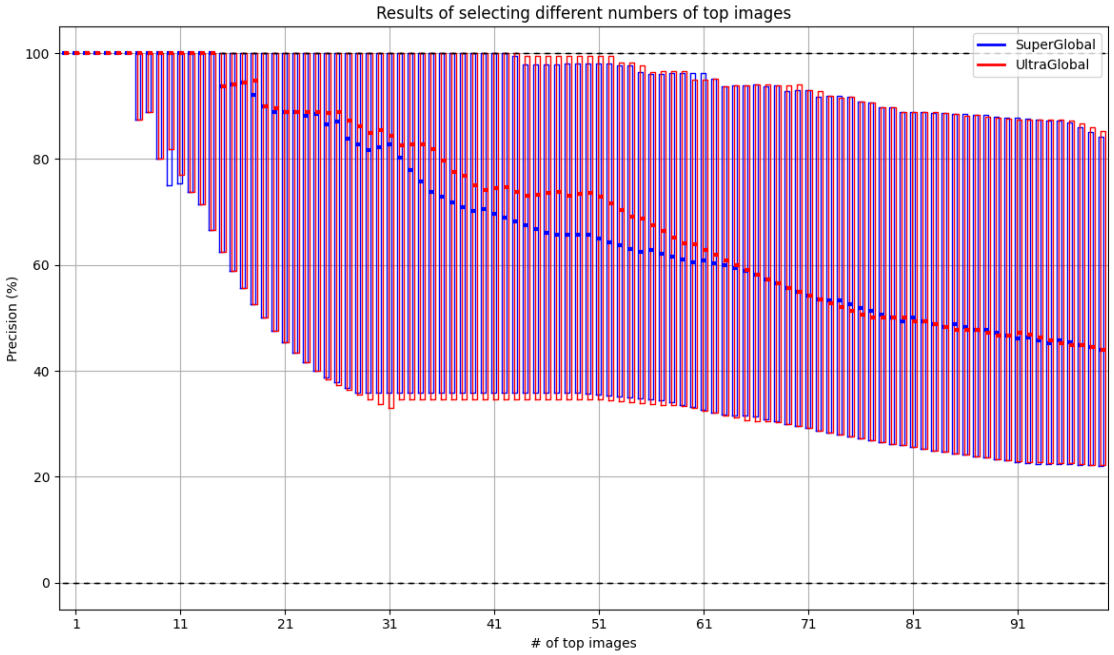

Fig. 4. Boxplots of precision for the top several results for all queries. Comparing UltraGlobal with SuperGlobal. Using ROxford(Hard) dataset.

To see the effect of different top results per query for the pre-improved SuperGlobal and the improved UltraGlobal, we made boxplots of precision for the top several results for all queries for both methods fig. 4. The horizontal axis represents the number of selected top results, and the vertical axis represents the AP of these selected results. The AP of all queries is summarized into a box. The upper and lower edges of the box indicates the two quartiles, respectively. And the median line of the box indicates the median. As can be seen, the quartiles are essentially the same for both, indicating that better and worse queries have essentially the same effect under both methods. However, in the top21 to top65 range, the median of our method is higher, which indicates that our method can give better results for general queries.

4.2.2 Robust Testing & Ablation Study of PANet

On the other hand, thanks to the use of PANet, UltraGlobal has improved the model's stability to a certain extent. As shown in table 2 & table 3, we fine-tuned the original model on a small toxic dataset, and it can be observed that UltraGlobal's performance falls between the original ResNet and the Toxic ResNet. The architecture with PANet achieved an average score that was 7.5% higher than the version without PANet integration.

Table 2: Comparison of toxic models (based on mAP metrics).

Model	With PANet	ResNet Depth	mAP(E)	mAP(M)	mAP(H)
Base Model50	False	50	1.74	2.47	1.08
Base Model101	False	50	1.75	3.12	2.38
Toxic Model50	False	50	1.59	2.37	1.06
PANet Toxic Model50	**True**	**50**	**1.81**	**2.44**	**0.99**
PANet Toxic Model101	**True**	**101**	**1.74**	**2.52**	**1.14**

Table 3: Comparison of toxic models (based on MPR metrics).

Model	MPR(E)	MPR(M)	MPR(H)
Base Model50	[2.94 2.94 2.94]	[2.86 3.43 3.43]	[0.00 0.57 0.71]
Base Model101	[2.94 4.41 3.82]	[4.29 5.14 5.00]	[1.43 2.00 2.57]
Toxic Model50	[4.41 2.94 2.65]	[4.29 3.14 3.14]	[0.00 0.29 0.81]
PANet Toxic Model50	**[8.82 4.12 3.53]**	**[10.00 4.57 4.14]**	**[1.43 0.57 0.86]**
PANet Toxic Model101	**[2.94 3.53 3.68]**	**[2.86 3.71 4.71]**	**[0.00 0.36 1.50]**

The experimental findings indicate that the integration of PANet results in enhanced stability and improvements in the mAP and MPR metrics in certain scenarios. After model contamination,

both mAP and MPR were significantly affected across tasks of varying difficulty; however, the addition of our PANet network effectively alleviated these issues. This demonstrates that our PANet network is capable of robustly extracting image features, even in complex tasks, showcasing its strong robustness and generalization ability.

We also conducted ablation studies on the toxic model to evaluate the impact of each component on overall model performance when PANet is either included or excluded. We choose the ResNet with a depth of 50 for these experiments. The results, as shown in the table 4 & table 5, indicate that using PANet for feature extraction yields superior performance compared to its absence when all other components are excluded, with the advantages becoming even more pronounced upon the inclusion of additional components.

Table 4: Comparison of the activation of different components (based on mAP metrics).

With PANet	ResNet Depth	Activate Component	mAP(E)	mAP(M)	mAP(H)
True	50	nan	1.70	2.46	1.09
True	50	gemp	1.94	3.21	2.37
True	50	sgem	1.69	3.13	2.41
True	50	regm	1.74	2.43	0.99
True	50	relup	1.88	2.57	1.11
True	50	rerank	1.73	2.40	0.99
False	50	nan	1.64	2.38	1.05
False	50	gemp	1.93	2.64	1.12
False	50	sgem	1.67	2.42	1.04
False	50	regm	1.55	2.99	2.37
False	50	relup	1.73	2.59	1.31
False	50	rerank	1.79	3.15	2.37

Table 5: Comparison of the activation of different components (based on MPR metrics).

With PANet	ResNet Depth	Activate Component	MPR(E)	MPR(M)	MPR(H)
True	50	nan	[5.88 3.24 3.68]	[7.14 4.00 4.29]	[1.43 1.14 1.14]
True	50	gemp	[8.82 4.41 2.65]	[10.00 5.14 4.00]	[1.43 2.00 2.71]
True	50	sgem	[2.94 2.65 2.50]	[4.29 4.00 3.86]	[1.43 2.57 2.86]
True	50	regm	[2.94 4.12 3.53]	[4.29 5.43 4.43]	[1.43 1.43 1.00]
True	50	relup	[10.29 3.82 4.41]	[10.00 4.57 5.14]	[0.00 1.14 1.14]
True	50	rerank	[5.88 5.29 4.41]	[7.14 6.00 5.14]	[1.43 0.86 0.86]
False	50	nan	[4.41 3.82 3.24]	[4.29 4.57 3.71]	[0.00 0.86 0.67]
False	50	gemp	[5.88 5.59 4.71]	[5.71 6.29 5.43]	[0.00 0.86 1.00]
False	50	sgem	[4.41 3.53 3.82]	[4.29 4.57 4.71]	[0.00 1.14 1.00]
False	50	regm	[1.47 2.06 2.94]	[2.86 2.57 3.57]	[1.43 1.71 2.00]
False	50	relup	[2.94 4.41 4.41]	[2.86 4.86 5.00]	[0.00 1.29 1.29]
False	50	rerank	[5.88 3.82 3.97]	[7.14 4.86 5.00]	[1.43 2.29 2.43]

Specifically, when gemp is enabled, the model demonstrates significant improvements across all categories, particularly with PANet, achieving mAP(M) and mAP(H) values of 3.21 and 2.37, re-

spectively. This suggests that gemp plays a beneficial role in enhancing performance. In the absence of PANet, both mAP and MPR are relatively satisfactory but exhibit slightly reduced effectiveness compared to configurations that include PANet.

The activation of sgem also led to improvements in both mAP and MPR metrics, with PANet markedly enhancing the model's feature extraction capabilities. Conversely, when rgem is enabled, the PAN structure yields good performance in mAP(E), although mAP(M) and mAP(H) values are lower. The MPR metrics in this case also displayed moderate performance. Notably, without PANet, regm outperformed expectations on Medium and Hard tasks, achieving mAP(M) and mAP(H) values of 2.99 and 2.37, respectively. This finding suggests that, in certain contexts, regm may exhibit superior performance in the absence of the PAN structure. Similarly, the rerank component demonstrated analogous performance patterns in Medium and Hard tasks.

When relup is activated within the PANet structure, it consistently performed well across all categories, showing significant improvements in mAP(E) and mAP(H). Its MPR metrics also indicated strong ranking performance in Easy and Medium tasks. Even without PANet, relup maintained a commendable level of performance, underscoring its stability as an activation component.

In summary, our comparisons reveal that the activation of gemp significantly enhances model capabilities, exhibiting superior performance irrespective of the presence of PANet. When considering both sgem and relup, the former shows superior performance within the PANet structure, while the latter demonstrates a more pronounced impact on model capability when PANet is not included. In addition, rgem appears to be better suited for models without PANet, and rerank can be selectively employed in specific task scenarios, leading to notable improvements in overall model performance.

5 Future Research Directions

5.1 Exploration of Model Lightweighting

While UltraGlobal effectively balances feature extraction quality and inference speed, an important future research direction involves further exploration of lightweight models to reduce computational overhead, thereby optimizing the system for image retrieval. Although the current model complexity impacts the feature extraction speed for query images, techniques such as model pruning and quantization could significantly reduce memory usage and computational demands without sacrificing accuracy. This would not only improve the system's response time but also enhance its applicability in resource-constrained environments. Optimizing UltraGlobal for calculation speed and memory usage holds great potential for expanding its use in real-time scenarios, particularly on mobile and embedded systems.

5.2 Enhancing Global Feature Extraction Methods

UltraGlobal relies predominantly on global feature extraction and encoding, and future research could improve the discriminative power of global descriptors by integrating more advanced attention mechanisms. For example, adding intermedia layers to the network architecture can improve the model's reasoning ability, which may improve the model's ability to find similar images; comprehensively consider local features and global features and integrate them to find features in more

receptive fields; introduce a U-net-like architecture, which can not only solve the overfitting problem, but also allow more pretrain models to be applied. These methods may be useful to improve the performance of the model to a certain extent, leaving room for future work.

5.3 Generalization to Diverse Scenarios and Datasets

While UltraGlobal demonstrates excellent performance across several benchmark datasets, further validation of its generalization capabilities in diverse scenarios and datasets is necessary. Future research could explore training and evaluation on larger and more visually diverse datasets to assess the model's adaptability and robustness in real-world applications. Introducing more varied data scenarios, such as low-light conditions, complex backgrounds, and different image styles, could enhance UltraGlobal's performance in diverse environments. Additionally, exploring self-supervised or unsupervised learning methods to fine-tune the model could reduce dependence on labeled datasets, thereby extending its applicability to domains where labeled data is scarce. This would significantly enhance UltraGlobal's practical utility, enabling efficient and accurate image retrieval in a broader range of real-world contexts.

6 Conclusion

In this paper, we presented UltraGlobal, an enhanced image retrieval method that addresses the key limitations of both traditional local-feature-based and deep learning approaches. By leveraging advanced techniques such as PANet for feature extraction, multiple improvements to GeMP pooling modules, and integrating VLAD encoding, UltraGlobal demonstrates superior scalability, efficiency, and precision. The method achieves state-of-the-art performance across standard benchmarks, significantly improving both global feature retrieval and reranking processes. Our experimental results on the ROxford and RParis datasets show that UltraGlobal outperforms existing methods, especially in challenging retrieval scenarios, demonstrating its robustness and effectiveness.

Looking ahead, the flexibility and robustness of UltraGlobal suggest promising potential for adaptation in various real-world applications, from large-scale digital asset management to real-time image search systems. Future improvements, particularly in model lightweighting and the use of more advanced attention mechanisms, could further optimize the method for edge computing and diverse, resource-constrained environments. Thus, UltraGlobal sets a new benchmark for efficient, scalable, and accurate image retrieval systems.

Acknowledgement

Xuanlang Dai, Shicheng Wang, Pengfei Huang, Zhiqi Zhang, and Mingyang Gao contributed equally to this work and should be considered co-first authors.

References

[1] Jégou, H., Douze, M., Schmid, C., & Pérez, P. (2010). Aggregating local descriptors into a compact image representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3304-3311.

[2] Radenović, F., Tolias, G., & Chum, O. (2018). Revisiting Oxford and Paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5167-5176.

[3] Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815-823.

[4] Tolias, G., Sicre, R., & Jégou, H. (2016). Particular object retrieval with integral max-pooling of CNN activations. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[5] Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., & Sivic, J. (2016). NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5297-5307.

[6] Gordo, A., Almazan, J., Revaud, J., & Larlus, D. (2017). End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision (IJCV)*, 124(2), 237-254.

[7] Radenović, F., Tolias, G., & Chum, O. (2018). Fine-tuning CNN image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(7), 1655-1668.

[8] Cao, B., Araujo, A., & Sim, J. (2020). Unifying deep local and global features for image search. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 726-743.

[9] Philbin, J., Chum, O., Isard, M., Sivic, J., & Zisserman, A. (2007). Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1-8.

[10] Avrithis, Y., & Tolias, G. (2014). Hough pyramid matching: Speeded-up geometry re-ranking for large scale image retrieval. *International Journal of Computer Vision (IJCV)*, 107(1), 1-19.

[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998-6008.

[12] Choy, C., Gwak, J., & Savarese, S. (2019). Fully convolutional geometric features. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 8958-8966.

[13] Shao, S., Chen, K., Karpur, A., Cui, Q., & Araújo, A. (2023). Global features are all you need for image retrieval and reranking. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

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

[15] CVNet: A Convolutional Neural Network for Image Recognition and Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[16] Dosovitskiy, A., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[17] Shao, S., Chen, K., Karpur, A., Cui, Q., & Araújo, A. (2023). Global features are all you need for image retrieval and reranking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[18] Lowe, D.G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision (ICCV)*, vol. 2, pp. 1150-1157, doi: 10.1109/ICCV.1999.790410.

[19] Lowe, D.G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2), 91-110. https://doi.org/10.1023/B:VISI.0000029664.99615.94.

[20] Csurka, G., & Perronnin, F. (2011). Fisher Vectors: Beyond Bag-of-Visual-Words Image Representations. In Richard, P., & Braz, J. (Eds.), *Computer Vision, Imaging and Computer Graphics. Theory and Applications. VISIGRAPP 2010*, Communications in Computer and Information Science, vol. 229, Springer, Berlin, Heidelberg, pp. 28-42. https://doi.org/10.1007/978-3-642-25382-9_2.

[21] Perronnin, F., Sánchez, J., & Mensink, T. (2010). Improving the Fisher Kernel for large-scale image classification. In Daniilidis, K., Maragos, P., & Paragios, N. (Eds.), *Computer Vision – ECCV 2010*, Lecture Notes in Computer Science, vol. 6314, Springer, Berlin, Heidelberg, pp. 143-156. https://doi.org/10.1007/978-3-642-15561-1_11.

[22] Jégou, H., Douze, M., Schmid, C., & Pérez, P. (2010). Aggregating local descriptors into a compact image representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3304-3311. doi: 10.1109/CVPR.2010.5540039.

[23] Tolias, G., Sicre, R., & Jégou, H. (2015). Particular object retrieval with integral max-pooling of CNN activations. *arXiv preprint arXiv:1511.05879*.

[24] Kalantidis, Y., Mellina, C., & Osindero, S. (2016). Cross-dimensional weighting for aggregated deep convolutional features. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part I*, Springer International Publishing, pp. 685-701.

[25] Noh, H., Araujo, A., Sim, J., Weyand, T., & Han, B. (2017). Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3456-3465.

[26] Shi, J., & Tomasi, C. (1994). Good features to track. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 593-600, doi: 10.1109/CVPR.1994.323794.

[27] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 30, pp. 5998-6008.

[28] Choy, C., Gwak, J., & Savarese, S. (2019). 4D spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3075-3084.

[29] Zhang, S., Liu, H., Lin, L., & Qiao, S. (2020). V4D: 4D convolutional neural networks for video-level representation learning. *arXiv preprint arXiv:2002.07442*.

[30] Wang, X., et al. (2018). Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7794-7803.

[31] T. Weyand, A. Araujo, B. Cao and J. Sim. Google Landmarks Dataset v2 – A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 2572-2581, doi: 10.1109/CVPR42600.2020.00265.

[32] M. Yang et al. DOLG: Single-Stage Image Retrieval with Deep Orthogonal Fusion of Local and Global Features. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 2021, pp. 11752-11761, doi: 10.1109/ICCV48922.2021.01156.