Machine Learning with Applications 15 (2024) 100523

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

SODRet: Instance retrieval using salient object detection for self-service shopping

Muhammad Umair Hassan [a,*], Xiuyang Zhao [b], Raheem Sarwar [c], Naif R. Aljohani [d], Ibrahim A. Hameed [a]

[a] Department of ICT and Natural Sciences, Norwegian University of Science and Technology (NTNU), Ålesund, Norway
[b] School of Information Science and Engineering, University of Jinan, China
[c] OTEHM, Manchester Metropolitan University, Manchester, United Kingdom
[d] Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Keywords:
Image retrieval
Salient object detection
Deep learning
Online shopping

ABSTRACT

Self-service shopping technologies have become commonplace in modern society. Although various innovative solutions have been adopted, there is still a gap in providing efficient services to consumers. Recent developments in mobile application technologies and internet-of-things devices promote information and knowledge dissemination by integrating innovative services to meet users' needs. We argue that object retrieval applications can be used to provide effective online or self-service shopping. Therefore, to fill this technological void, this study aims to propose an object retrieval system using a fusion-based salient object detection (SOD) method. The SOD has attracted significant attention, and recently many heuristic computational models have been developed for object detection. It has been widely used in object detection and retrieval applications. This work proposes an instance retrieval system based on the SOD to find the objects from the commodity datasets. A prediction about the object's position is made using the saliency detection system through a saliency model, and the proposed SOD-based retrieval (SODRet) framework uses saliency maps for retrieving the searched items. The method proposed in this work is evaluated on INSTRE and Flickr32 datasets. Our proposed work outperforms state-of-the-art object retrieval methods and can further be employed for large-scale self-service shopping-based points of sales.

1. Introduction

Technological innovation is the main driver for economic growth and human progress (Zhang, Yang, & Yang, 2022). With the swift development of online media and associated algorithms (Hechavarria & Shafiq, 2022; Ozbay & Alatas, 2021), especially the popularity of social networking sites such as Facebook, Flickr, Weibo, and online trading platforms such as Amazon and Taobao (Cauteruccio, Corradini, Terracina, Ursino, & Virgili, 2022; Khanam, Srivastava, & Mago, 2022), various image data proliferate every second (Nasirtafreshi, 2022). For example, there were 248 million daily active users on Weibo in 2021 (Thomala, 2021), more than 100 million images are uploaded daily, and tens of billions of images are kept on Taobao and JD.com's servers. Given the rich and massive number of digital images, quickly and accurately finding the images that users want to see in the image library has become a core research area in information retrieval. Steering on the new trends of mobile applications technology, enterprises are developing communication platforms for the dissemination of knowledge, social media integration, and online shopping (Hsu & Tang, 2020). Thanks to large-scale online data, a rapid boom in technological innovation has been achieved. It is of great interest to show customers their intended products while searching online, which also helps firms satisfy consumers' needs (Guo & Lv, 2022; Pirnay & Burnay, 2022).

Image retrieval has always been a research hotspot in information processing, and researchers have continuously explored this aspect. Image retrieval technology is not only used in image search engines but also widely used in computer vision applications (Dost, Serafini, Rospocher, Ballan, & Sperduti, 2022; Kaur & Singh, 2022). In the middle of the last century, image retrieval research began—the original method used image keywords to search for images. Primarily, text-based image retrieval avoids image content recognition to a certain extent. It fits the user's well-known retrieval habits and is relatively simple in system implementation, but its disadvantages are no longer invisible. The first problem is that the textual description is sometimes

* Corresponding author.
E-mail addresses: muhammad.u.hassan@ntnu.no (M.U. Hassan), zhaoxy@ujn.edu.cn (X. Zhao), r.sarwar@mmu.ac.uk (R. Sarwar), nraljohani@kau.edu.sa (N.R. Aljohani), ibib@ntnu.no (I.A. Hameed).

https://doi.org/10.1016/j.mlwa.2023.100523
Received 23 August 2023; Received in revised form 26 November 2023; Accepted 18 December 2023
Available online 29 December 2023

confusing in expressing the query image's visual information entirely. Secondly, the text content and keywords of artificially added images are subjective and lack uniform standards, which can cause retrieval errors. Hence, content-based image retrieval (CBIR) methods are developed (Gudivada & Raghavan, 1995; Kumar et al., 2022; Wang, Jiao, Liu, Ma, & Shang, 2022).

CBIR's algorithmic process analyses the image content and then extracts a series of visual information, such as colour, shape, and texture, and combines them into the image. The descriptors of the image are stored in the feature library. Furthermore, the retrieval process works by extracting the image descriptors for a given query image. Then, some similarity matching algorithms to measure the similarity between the query image and the images in the image library, and finally, according to the level of similarity, the image library and the query images with high similarity are sent back to the user (Wang et al., 2020).

Recently, remarkable progress has been observed in multiple self-service online and E-commerce shopping stores such as Amazon, Alibaba (Zhang et al., 2018), JingDong, and TaoBao. Precisely retrieving commodities from the datasets is hard to manage for these online shopping applications (Fang et al., 2016; Ji, Zhang, Zhang, & Liu, 2021). This work proposes an algorithm to retrieve the products effectively and precisely to simplify the object retrieval process. We propose a salient object detection system that can find the rough location of a commodity item. Usually, the images of the commodities are pretty visible and adequately fit for image search. Along with the object's rough location, we also tend to see many useless proposals that prompt us to develop a feature that can help decrease the number of proposals. Therefore, we have developed a saliency system to perform object retrieval in two stages. First, it will find the approximated location of the object, that is, the salient object detection part of this work. Then, a result will be filtered again through the saliency box to remove unnecessary proposals. These two steps are performed simultaneously and help find the object searched selectively with more precision and accuracy by retaining the queried object and eliminating the unrelated candidate proposals.

During the initial stage, the images containing the objects mentioned in the search are searched. The objects are temporarily outlined or highlighted using a bounding box in the second step. The main focus of our proposed framework is to improve the object retrieval process by locating the object more accurately and precisely and reducing the number of other proposals and recommendations. The first step consists of salient object detection, whereas in the second step, we perform object retrieval for the selected images to be searched. The summary of contributions we made in this work is as follows:

- We proposed a saliency detection-based retrieval method (SODRet) in which the saliency map is generated by using a CNN-based fusion of the salient object detection (SOD) methods. In our improved work, the saliency box can accurately and precisely find the location of objects that need to be retrieved.
- The number of candidate proposals can be decreased by filtering out the proposals obtained by selective search (Cordonnier et al., 2021) by using a saliency box on the image patches. We can quickly retrieve objects and take less time to match the candidate proposals with this.
- In this work, we represent a combination of benefits from SOD and channel weighting generalized mean pooling (CGWMP) (Wang, Liu, et al., 2018) resulting in feature representation. This process will prevent intermittent transmission in convolution channels.

To assess the performance of the proposed SODRet framework, we performed extensive experiments on INSTRE (Wang & Jiang, 2015) and Flickr32 (Romberg, Pueyo, Lienhart, & Van Zwol, 2011) datasets. The datasets are selected based on a careful examination of the literature on object retrieval methodologies. We compared SODRet with traditional and deep learning-based retrieval approaches. For a fair comparison, we adopted mean average precision (mAP) for all works reported in this study. From the experiments, we show that our SODRet outperforms all state-of-the-art methods.

Following is the organization of our proposed work. Section 2 delivers a thorough overview of related object retrieval and deep learning works, for instance-level image retrieval. The methodology of our proposed SODRet is available in Section 3. The experimental results are shown in Section 4. We have briefly discussed this work in Section 5 while the concluding remarks are given in Section 6.

2. Related work

These days, many commodity items are searched from large-scale search engines, for example, Google Image Search, TinEye, Bing Image Feed, etc. Many popular e-commerce stores, e.g., Amazon and eBay, integrated the instance-level search to help users find their favourite products (Zhang et al., 2018). The process of retrieving products out of the datasets, precisely, is a hard job to manage for these online shopping applications.

Over the last few decades, there have been remarkable achievements in object retrieval and recognition. Real-world objects are recognized and organized based on human cognition and perceptual ability to find the similarity among object classes. Based on the object of interest's perceptual ability, the object retrieval system can be divided into two major retrieval classes: instance level and category level. For example, at the instance-level, "the Starbucks bottle" and "the Coca-Cola bottle" are all instance-level labels that are the subordinate class of category level "bottle". The instance-level visual task is of great importance and has applicable values.

According to previous works, the images of objects are dependent on single local invariant descriptors; one of them is scale-invariant feature transform (SIFT) (Lowe, 1999). Bag-of-words (BOW) (Sivic & Zisserman, 2003) is another commonly used method that uses a bag of local descriptors to find a small number of highly distinctive features of image representation. Babenko and Lempitsky (2015) showed that the full connection layer features are appropriate for retrieving the object image. The connection layer performs the classification of images as they are trained to detect objects' labels. This feature is not found in other types and methods. Another work by Tolias and Jégou (2014) performed the aggregation of regional-maximum activation of convolutional layers (R-MAC), which resulted in compact representation, and their approach outperformed all the other representations. Chen, Kuang, Wong, and Zhang (2017) presented another feature in which activations of convolution are clustered and then combined most of the initiations from these collections to reduce the feature limitation of R-MAC. Mohedano, McGuinness, Giró-i Nieto, and O'Connor (2018) applied the saliency concept based on bags-of-local convolutional features (BLCF) for object retrieval.

2.1. Object location

According to the previous studies, there are two diverse categories of methods for locating the objects. The first type is a deep neural network-based end-to-end supervised learning of object retrieval. The network can learn the location and labels of the objects simultaneously. A relative ranking of the images is obtained by taking advantage of a Regional Proposal Network (RPN) (Salvador, Giró-i Nieto, Marqués, & Satoh, 2016). The RPN can reduce the computation time of producing advanced-quality region proposals by utilizing multiple anchor scales. The second type of object locating method is unsupervised, contrary to the first method (Viola & Jones, 2004). The process of reserving the bounding boxes in the EdgeBox (Zitnick & Dollár, 2014) method consists of measuring the number of edges in the candidate box and subtracting this from the number of edges in the other overlapping boundary boxes, and finding the upper box objective score. In the selective search method (Uijlings, Van De Sande, Gevers, & Smeulders,

2013), hierarchical segmentation is combined with a thorough search. This method uses various hierarchical and complementary grouping strategies to find and filter the quality of the advanced object proposals, which are also independent of the class. Selective search (Uijlings et al., 2013) and EdgeBox (Zitnick & Dollár, 2014) methods can help reduce the search space compared with the previously used brute force searching methods.

2.2. Neural networks

CNNs are known as conventional methods applicable in various tasks successfully—for example, retrieval of images, classification of images, and detecting objects. Out of a large amount of labelled data, CNN can take out more affluent semantic information because of its deep learning architecture compared to traditional and commonly used visual methods. Compared to image classification, extensive research is done on descriptors based on CNN; however, it is relatively less for instance/object retrieval. Razavian, Sullivan, Carlsson, and Maki (2016) used several tasks as test tasks, including instance retrieval, which is used to assess the CNN model's performance. For image retrieval on a comparatively more extensive scale, Chandrasekhar, Lin, Morère, Veillard, and Goh (2015) changed the sparse high-dimensional CNN representation to very compressed representations and called it the hash method. These works have similar research areas as proposed in our work and show how the pre-trained CNN used its convolution features. Still, we decided to choose the object detection CNN for instance-level image retrieval, a state-of-the-art method for extracting region-based convolution features.

Detection pipelines of CNN-based objects have been anticipated in many works previously done. Girshick, Donahue, Darrell, and Malik (2014) presented the R-CNN, in which, as an input, the object region proposals were used in the network instead of full images. Fully connected layers were extracted for all windows at test time, and training was provided with a regressor and classifier for the bounding box. Sharif Razavian, Azizpour, Sullivan, and Carlsson (2014) presented Faster R-CNN, which helped eliminate the reliance of former CNN systems for object detection on object proposal by proposing a Region Proposal Network. Some works have been applied to detect an object, such as YOLO v2 (Sang et al., 2018) and single shot detector (SSD) (Liu et al., 2016). In addition to the above, we used various features and infused strategies to top-ranked detected regions to compare the results of state-of-the-art methods.

2.3. Instance level image retrieval

There are two major blocks of the state-of-the-art instance-level image retrieval pipelines (Revaud, Almazán, Rezende, & Souza, 2019). The first block is a subset of object images retrieved from a database, similar to the query image. The relevant images with high precision are selected from the subset by applying geometric consistency checks. In the first step, high-dimensional vectors are compared against tens of thousands of dimensions, which can represent the image's content. Using better global descriptors is the key to improving retrieval performance. The multimedia research community has been paying more attention to some applications (Wray, Larlus, Csurka, & Damen, 2019), such as visual search, digital documents, dispersed large-scale search, and compressed descriptors for real-world applications. Attention-based image retrieval methods have been proven to achieve more precise results, according to recent studies (Chaudhuri, Banerjee, Bhattacharya, & Datcu, 2020). Other studies have gone ahead of the instance-level retrieval, aiming to find the image having the same semantics as the query image.

In an attempt, Bhunia et al. (2023) proposed a unique approach that focused on how hand-drawn drawings might be used to describe "salient object". To do this, they provided a photo-to-sketch generation model that uses a 2D attention mechanism to produce sequential drawing coordinates matching a given visual photo. Salient areas in the process are created by attention maps that are collected over the course of the time steps. In another effort, a Similarity Retrieval-based Inference Network (SRI-Net) is suggested by Lv et al. (2023). Because different focus points provide different focused slices from light field pictures that are useful for salient object recognition, the main feature of this model is to choose the most useful focal slice that can provide more complementary information for the RGB image.

3. Salient object detection based retrieval

We are aware that the market products' images are marked and appropriately labelled and are clearly visible; therefore, the objects can be located quickly by a salient object detection system using the simulating human visual attention system. We remove the pixel-wise object class segmentation to build a simplified salient object detection network and refer to a multitask, fully convolutional neural network. Fig. 1 is the high-level illustration of the proposed SODRet. The input image is sent to the SOD network, where a saliency map is generated, and using the selective search of image patches, a saliency bounding box is predicted, which is further fed to a CNN to obtain a final ranking of retrieved images.

In the proposed network, a saliency map is generated initially; first, we apply the dense up-sampling of the input image for super-pixels segmentation and convert the input image into super-pixels (as can be seen in Fig. 2). Using the CNN-based fusion method, we fuse two saliency maps to get a final map for commodity objects. The designed network can select the most relevant object from the query image. The CWGMP (Wang, Liu, et al., 2018) strategy is applied for object retrieval, which preserves the distinctiveness of the convolution feature and helps achieve the burstiness or the intermittent transmission of data in the convolution channels effectively and efficiently.

The model has a VGG16 network. After feeding the image to the network, the convolution size is 3×3, with 128 feature maps. After this, we apply a max-pooling layer, making the convolution size 1×1 by retaining the feature channels. The significant feature extraction makes the channel size 1 and convolution size 1×1. A fusion-based CNN model is used for the salient object detection task. We used hypergraph-based object detection (Zhang, Wang, Lv, & Zhang, 2021) with the aggregation-based multi-level feature extraction (Zhang, Wang, Lu, Wang, & Ruan, 2017) method to generate a final saliency map using the feature integration of both methods. Our proposed work obtains a saliency map to feed it to the retrieval network. The framework of the pre-trained model is illustrated in Fig. 3. A set of training images in our network is represented by $X = \{x_i\}_{i=1}^{N}$, and corresponding to that, the ground truth-binary training maps of salient objects is represented is denoted by, $\{x_i\}_{i=1}^{N}$, All the parameter in the network are defined by θ. The saliency detection function is d. With the regularizing of all the training samples, the loss function L (see Eq. (1)) is minimized by using the gaze estimation of gradients (Liu et al., 2021) for the training purpose of the network.

$$L(X,\theta) = \frac{1}{N}\sum_{i=1}^{N}\left|X_i - d\left(x_i,\theta\right)\right| \tag{1}$$

The experiment results have shown that the training models can be applied to different datasets because they are highly generalized. Fig. 4 shows some of the saliency detection results.

3.1. Selecting saliency

Fig. 4 contains some impurities and blurred edges of objects in the saliency map generated from the saliency object detection network. These problems are dealt with SODRet using the proposed algorithm by Wang, Liu, et al. (2018). Algorithm 1 refines optimizing the saliency proposal generation and the saliency map. Following are the steps for refining using the proposed algorithm.

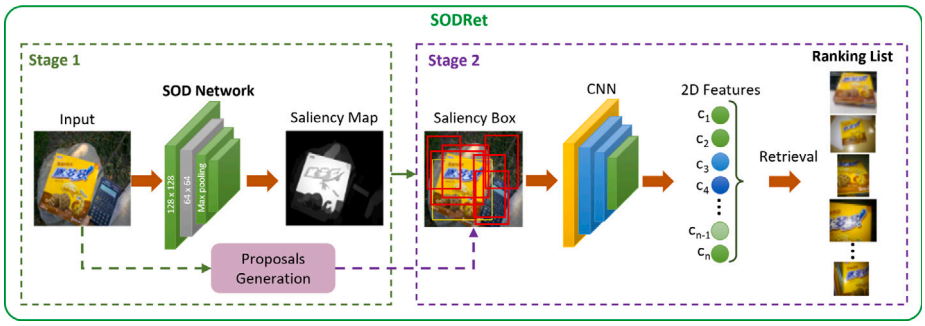


Fig. 1. The framework of object retrieval based on salient object detection — SODRet. In stage 1 of the proposed SODRet framework, an SOD mechanism is applied, as proposals for the commodity items are sent to stage 2. In stage 2, a saliency box generates the proposals for commodity items. A CNN-based retrieval framework is applied, and 2D features are generated. After that, a final ranked list for the retrieved items is generated.

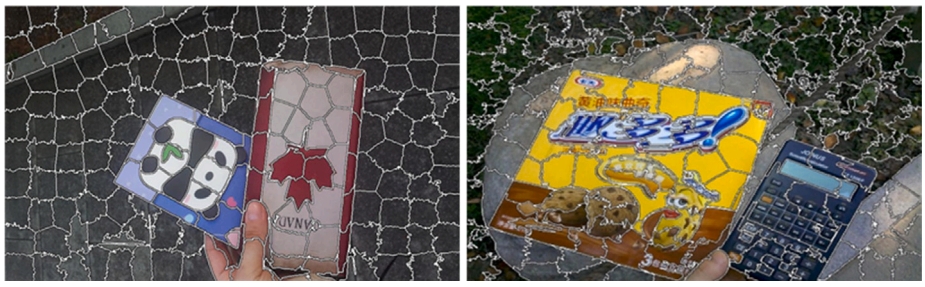


Fig. 2. The superpixels generated using the methodology described in (Wang, Chen, et al., 2018) as integrated into our salient object detection framework. These superpixels represent an over-segmentation of the input images into perceptually meaningful, homogeneous regions, which are fundamental to our model's ability to discern and highlight salient regions accurately.

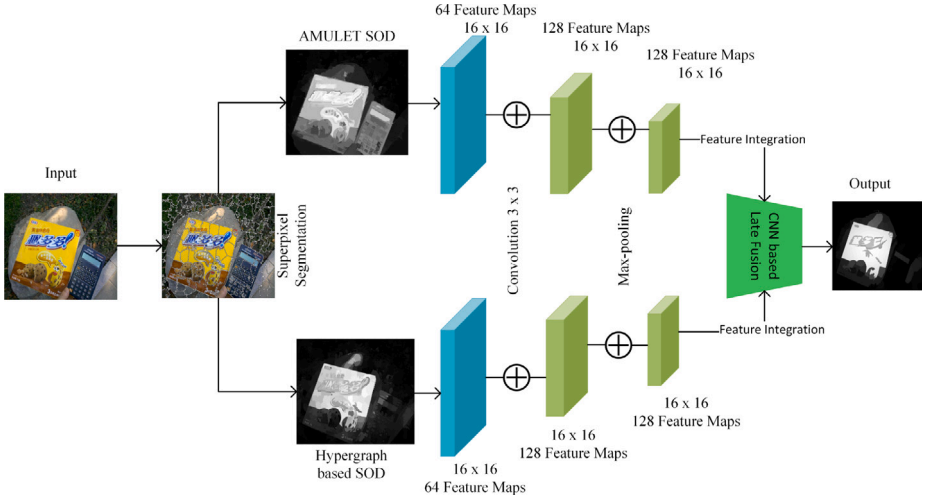


Fig. 3. The architecture of our advanced salient object detection network, specifically engineered for the instance retrieval task. This network employs a novel fusion strategy combining the novel Hypergraph-based saliency detection and AMULET, a state-of-the-art methods. At the core of this integration is a CNN-fusion model that meticulously merges saliency maps produced by both methods.

Fig. 4. The results of saliency detection achieved through our proposed method. The result is a highly refined saliency map that accurately represents the prominent objects in the scene and delineates them with precise boundaries.

Algorithm 1 Proposal Selection Process

```
1: procedure ProposalSelection
2:     calculate saliency map   ▷ Identify prominent areas in the image
3:     calculate pixel value    ▷ Determine the value of each pixel
4:     for each pixel in image do
5:         if pixel value < threshold then
6:             pixel value ← 0
7:         else
8:             pixel value ← 255
9:         end if
10:    end for
11:    MUR (Maximum Unicom Region) ← Empty
12:    MBR (Minimum Bounding Rectangle) ← Empty
13:    calculate MUR based on thresholded image
14:    MBR ← Calculate Minimum Bounding Rectangle of MUR
15:    Filtering proposals:
16:    for each proposal region do
17:        calculate IoU (Intersection over Union)
18:        calculate IoR (Intersection over Region)
19:        if IoU or IoR meets criteria then
20:            Select proposal
21:        end if
22:    end for
23:    return selected proposals
24: end procedure
```

In the Algorithm 1, MUR on line 11 represents the maximum uniform region used to calculate the area of the marked region for binary images. On the other hand, MBR stands for minimum bounding rectangle, which is used as a saliency box. An average pixel value of the saliency map is calculated. If the pixel value is less than the average, it is adjusted to 0; in other cases, it is 255. This ensures the transformation of a saliency map into images in binary form. Another issue is optimization, which we resolved using the gaze estimation of gradients with our saliency map generation algorithm.

We explain the Algorithm 1 as follows. Line 2 calculates the saliency map; on line 3, a pixel value is calculated, on line 4, we see if the pixel value is less than the average. If it is, the pixel value is set to 0. This thresholding step binarizes the saliency map, turning less salient pixels to 0. On line 5, if the pixel value is not less than the average, it is set to 255. This completes the binarization — more salient pixels are set to maximum.

The overlapping ratio between the proposal taken out from the selective search of image patches and the saliency box is calculated to find out the number of redundant proposals in order to filter them out. A threshold of the overlapping ratio is defined if the ratio of the proposal is greater than the threshold values. The saliency proposals are the candidate proposals, a collection of saliency boxes, and preserved proposals.

The usual method of evaluating the overlapping ratio in the filtering stage is an intersection over union function (IoU) (Rahman & Wang, 2016), which is calculated by the following equation:

$$IoU = \frac{PA \cap SA}{PA \cup SA} \tag{2}$$

In Eq. (2), PA is the proposal area while SA is the saliency area. This calculation shows a larger overlapping ratio for the region results only for a similar area and position as of the saliency box. So, the method of calculating the overlapping ratio is amended. The equation is now named as intersection over region ratio (IoR), as follows:

$$IoR = \frac{PA \cap SA}{PA} \tag{3}$$

If the images have complicated backgrounds, the size of the saliency box is more significant. When we use the normal IoU to calculate the overlapping ratio, if the threshold of the overlapping ratio is set larger, the smaller proposals' intersection with the saliency box will be lost easily. These proposals can be retained easily using the IoR calculation formula for overlapping ratios.

Recalling the object that needs retrieving is ensured. An example of this process is shown in Fig. 5 . A theoretical framework of the stage of filtering is shown in Fig. 5**(a)**. When using the IoU overlapping ratio formula, the bounding boxes, shown in purple, yellow, and blue colours, will be eliminated if the threshold setting is higher. Additionally, all the proposal and saliency boxes that have a relationship with each other will be retained. The real images are shown as an example in Fig. 5**(b)** using the IoU and IoR methods to evaluate the overlapping ratios, respectively. If the overlapping ratio has a threshold adjusted to 0.5 for both IoU and IoR, the latter will locate the object accurately out of all the reserved proposals. On the other hand, with the same threshold, if IoU is used, the reserved proposals and saliency boxes have similar locations, which can be disastrous for complex images. Fig. 5 textbf(c) shows the saliency objects which are detected for retrieval purposes.

3.2. Object retrieval based feature representation

This section provides information about the feature representation of objects obtained from salient object-based proposals. The saliency proposals are obtained once the proposal generation process is completed. For example, consider an image I having a size of $W \times H$, we form a 3D tensor of input image through the convolution layer having the size $W_1 \times H_1 \times F_1$, where F represents the output feature channels count. The architecture of the network and the resolution of the input image have an impact on the resolution $W_1 \times H_1$ of the 3D tensor.

The ReLU function activates every layer of the convolution; therefore, the elements of the feature map are either zero or positive. This 3D tensor can also be seen as 2D feature channel set responses: $C = \{C_i\}, i = 1, \ldots, K$, where C_i is the 2D tensor representing the responses of i_m^{th} feature channel. The convolution responses $C = \{C_i^R\}, i = 1, \ldots, K$ corresponding to saliency proposals R are obtained as per the network's scaling ratio. Generally, the convolution responses are handled by average pooling operation or the max pooling. These operations take input C_i^R, and a vector f is produced as an output of pooling processes. Every neighbour input selects only one node; therefore, the relevance of the regional feature is ignored by the max pooling operation, and the distinctiveness of the convolution feature is retained. Additionally, the average pooling operation preserved the correlations between regional features. However, the individual node contribution is attenuated and enhanced, ignoring the need for local structures. The method is as follows:

$$f = \left[f_1, \ldots, f_i, \ldots, f_F\right]^H \tag{4}$$

$$f_i = r_i \left(\frac{1}{W_i^R} \sum_{x \in W_i^R} x^a \right)^{1/a} \tag{5}$$

In Eq. (5) r_i is the i^{th} channel; weight α is an experimental parameter where $\alpha \to \infty, f_i \to r_i \max W_i^R$ and when $\alpha = 1$. We examine a channel-wise feature aggregation process with a parametrized pooling operation in Eq. (5). Where α is a parameter that adjusts the pooling operation. As $\alpha \to \infty$, the operation approximates max pooling, emphasizing the strongest activations. When $\alpha = 1$, the operation equals average pooling, considering all activations equally. Since we have used the context of generalized mean pooling, when $\alpha = 1$, the generalized mean (which in this case is the pooling operation) becomes the arithmetic mean. This is because raising each element to the power of 1 does not change the value of the elements, and taking the first root of the average is just the average itself.

The combined feature is derived using all activation responses on a channel, the channel on which the features are repeated frequently,

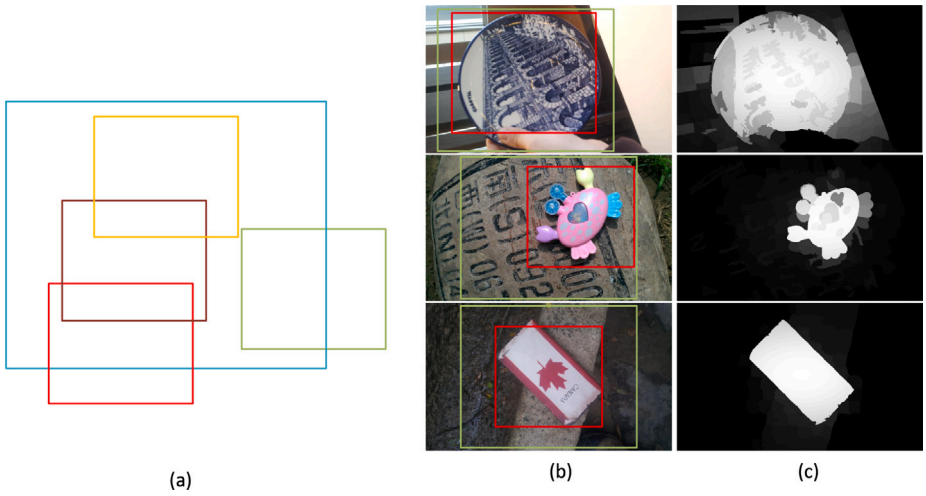


Fig. 5. Proposals generation using saliency maps. **(a)** The bounding boxes for selected items to be searched from the database. The red bounding box shows our selective item in **(b)**, and a saliency map is generated in **(c)**. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and the combined feature is activated strongly. If the features infrequently occur, the information provided could be distinct. Therefore, channel weighting, which uses sparsity-sensitive channel weighting, is produced. This channel matches the inverse document frequency. We can enhance and make the most out of the role of the channel with features that occur less frequently in the accumulated feature. The channel weight r_i is defined using the following:

$$r_i = \log\left(\frac{\sum_{i=1}^{H}\left(\frac{1}{G^R H^R}\sum_{x\in W_i^R} x^a\right)}{\in + \frac{1}{G^R H^R}\sum_{x\in W_i^R} x^a}\right) \tag{6}$$

Where $1 \in \{0,1\}$ is an indicator; $\in$ is a small constant added for numerical stability; G^R and H^R are the width and height of W^R, respectively. The feature vector dimensions are equal to H, which comprise a single value per channel. The image feature representation is the l_2-normalized vector f.

The feature maps in this work are produced by a pre-trained CNN when the fully connected layers are absent. Exhaustive Euclidean search is done for image retrieval according to the query feature over the database features. We used Euclidean distance to measure the image similarity. In order to get the preliminary rank list, the highest score of similarity of retrievals is used, which is viewed as the similarity scale of images that correspond to the query image. For retrieving the object in the image, the proposal having the highest similarity score is used to find the object's location. The resemblance between the normal vector and the top images' proposal features is calculated again to gain the final rank list.

4. Experimental evaluation

This section briefly describes the datasets and evaluation metrics used to perform the extensive experimentation of our work.

4.1. Datasets

- **INSTRE Dataset:** The INSTRE dataset contains three subsets called INSTRE S1, INSTRE S2, and INSTRE M. The number of images in each subset is 11,011 with 100 classes with single labels in INSTRE S1, and 12,059 images with 100 classes of single labels in INSTRE S2. INSTRE M contains the images of the objects that are also part of the other two subsets. We used the Caffe library to implement the saliency object detection method for our object retrieval work.
- **Flickr32 Dataset:** The Flickr32 dataset contains 8240 images collected from 32 logo brands. The images in logos have estimated planes.

4.2. Instance retrieval performance evaluation

This section examines and compares multiple aspects of our method with relevant image retrieval methods. We used INSTRE (Wang & Jiang, 2015) and Flickr32 (Romberg et al., 2011) datasets during the experiments to evaluate the performance of our proposed method. Fig. 6 represents the comparison of our method with other state-of-the-art approaches. It can be seen that the saliency maps generated by our SOD method yield in accurately filtering the proposals.

In order to train the network, we used 5000 images in INSTRE and another 1232 images used as a test set. In order to prepare the initial 13 layers of this network, we use a pre-trained VGG16 network, and to initialize the deconvolution, we use the simple bilinear interpolation. The momentum parameter is set to 0.96, 0.00002 is the learning rate, 0.0003 is the weight decay, and the 40,000 is the maximum iteration. Five images from each category are randomly chosen as query images from INSTRE and Flickr32 datasets; this results in 1250 query images in the INSTRE and 160 images in the Flickr32 datasets.

4.3. Mean average precision

Object retrieval performance can be calculated in relation to mean average precision (mAP). Image retrieval performance can be calculated using the following:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} p_i \frac{\mathrm{Pr}_i}{N_p} \tag{7}$$

Where N is the number of images in the dataset. p_i is a binary function. If the i^{th} position in the rank list is the correct result, $p_i = 1$, otherwise 0. Pr_i indicates the precision of the first i retrieval results. N_p indicate the number of query images which are related by the number of images.

The standard protocols were followed in all experiments. As input to the pre-trained VGG16 network, the query images are cropped with the bounding boxes. The convolutional layer has 512 feature channels for VGG16; we used the last convolution layer to extract our representation. As per the experiments, we have set the overlapping ratio threshold to 0.5, and the value of our feature map α is set to 4 in the remaining work.

The experimental results of different pooling methods are shown in Fig. 7**(a)**, which is in contrast with varying methods of pooling. In Fig. 7**(a)**, SB stands for saliency box; SP is the saliency proposal; AP is average pooling; GMP denotes unweighted generalized mean pooling. The unit for the mAP is the percentage. Our proposed feature

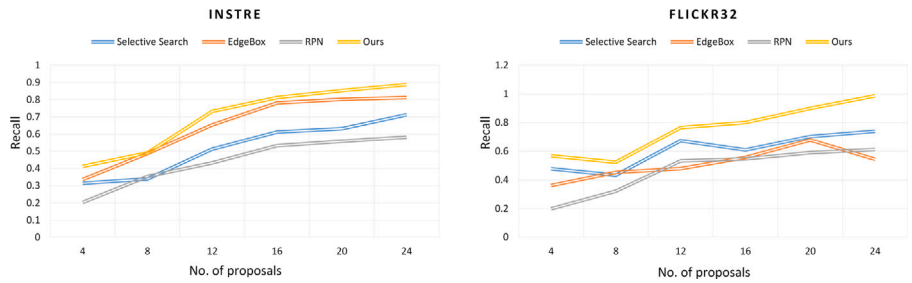
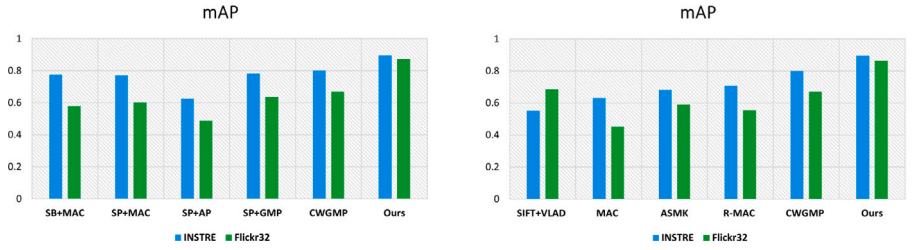


Fig. 6. Average recall of our proposed salient object detection framework. The abscissa shows the number of proposals used in our network as well as ordinate shows average recall on both datasets, respectively.



(a) mAP for pooling-based methods. (b) mAP for SIFT and CNN-based methods.

Fig. 7. The mAP (%) comparison of the proposed method with pooling-based methods in **(a)**, and with SIFT and CNN-based methods in **(b)** on two datasets.

representation has a higher performance rate than the other features, such as the unweighted generalized mean pooling feature, the average pooling feature, and the MAC feature (Kim & Yoon, 2018). Our feature preserves the discriminability and correlation of convolution features by aggregating the advantages of max pooling and average pooling. Our method takes the sparsity of the convolutional channel into account. This helps suppress the intermittent transmission of data into the convolution channel.

In addition to that, we divided the methods into two types in Fig. 7**(b)**: SIFT-based methods and CNN-based methods. The unit for mAP is the percentage; we have compared our method with most related methods in detail. The methods that we used for the comparison are SIFT and VLAD based method (Jégou, Douze, Schmid, & Pérez, 2010), MAC (Sharif Razavian, Sullivan, Maki, & Carlsson, 2015), ASMK (Tolias, Avrithis, & Jégou, 2016), R-MAC (Tolias, Sicre, & Jégou, 2015), and CWGMP (Wang, Liu, et al., 2018). The mAP results yielded by our method on INSTRE and Flickr32 are 82.0 and 69.2, respectively; these values ensure the best performance out of all other methods.

4.4. F1-score

The F1-Score is a way to balance the precision and recall of a system. It is instrumental in scenarios like image retrieval, where we want to understand how effectively your system retrieves relevant images while minimizing irrelevant ones. We calculated F1-Score to evaluate how effectively our proposed work retrieves relevant images from the commodity items framework. We also compared the retrieved instances with other state-of-the-art works. To calculate the F1-score, we used the following formulation.

Precision: This is the proportion of relevant retrieved images and is calculated as follows.

$$\text{Precision} = \frac{\text{Number of relevant images retrieved}}{\text{Total number of images retrieved}} \tag{8}$$

Recall: This measures the proportion of relevant images retrieved from all relevant images available and formulated as follows.

$$\text{Recall} = \frac{\text{Number of relevant images retrieved}}{\text{Total number of relevant images available}} \tag{9}$$

The final F1-Score for retrieved images is calculated as:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

Fig. 8 shows the state-of-the-art performance of our proposed method on precision Fig. 8(a), recall Fig. 8(b), accuracy Fig. 8(c), and F1-score Fig. 8(d) metrics. The performance is achieved for training and testing data on 1, 3, 5, and 10-fold cross-validations across INSTRE and Flickr32 datasets.

In Fig. 8, the training and testing performance of our proposed model is available, and it can be seen that the SODRet effectively achieved state-of-the-art performance for the logos retrieved for both datasets. For the INSTRE dataset, the model's performance has been significantly better across all cross-validation scenarios. However, compared to INSTRE, the performance is a little worse for the Flick32 dataset because Flickr32 is a smaller dataset containing a limited number of images per class, and the class imbalance problem occurs for uneven distribution of samples across its 32 classes.

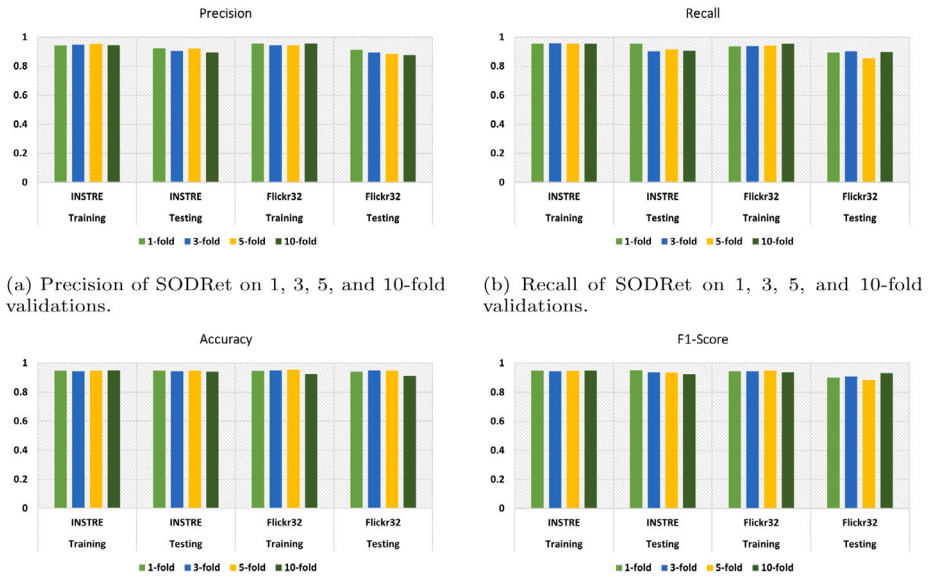
4.5. ROC/AUC

The Receiver Operating Characteristic (ROC) curve is a graphical abstraction to illustrate the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The Area Under the ROC Curve (AUC) provides an aggregate performance measure across all possible classification thresholds. It is particularly useful for evaluating the performance across all classes. To evaluate the performance across the five-fold cross-validation of our proposed model, we performed the experiments on both INSTRE and Flick32 datasets. The formulation of ROC curve is plotted with the True Positive Rate (TPR) against the False Positive Rate (FPR) where TPR is on the y-axis and FPR is on the x-axis. The equation for TPR and FPR is given below.

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{12}$$

Fig. 9 shows the ROC/AUC curves for INSTRE in Fig. 9(a) and Flick32 in Fig. 9(b) datasets. The findings indicate that our proposed model achieved 94%, 95%, 95%, and 94% mean AUC on 1, 3, 5, and



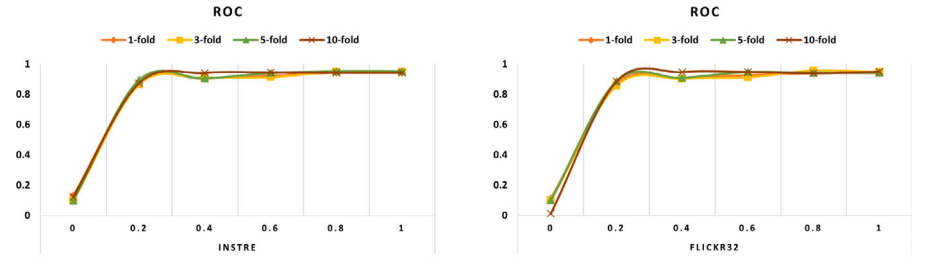
(a) Precision of SODRet on 1, 3, 5, and 10-fold validations.

(b) Recall of SODRet on 1, 3, 5, and 10-fold validations.

(c) Accuracy of SODRet on 1, 3, 5, and 10-fold validations.

(d) F1-Score of SODRet on 1, 3, 5, and 10-fold validations.

Fig. 8. Results for precision, recall, accuracy, and F1-score achieved for training and testing scenarios for 1, 3, 5, and 10-fold cross-validations on INSTRE and Flickr32 datasets.



(a) ROC curve on INSTRE dataset.

(b) ROC curve on Flickr32 dataset.

Fig. 9. The ROC curve of proposed SODRet generated for INSTRE and Flickr32 datasets for four-fold cross-validations.

10-fold cross-validations on the INSTRE dataset, respectively, whereas, on the Flickr32 dataset, the mean AUC is 94%, 94%, 94%, and 95% for 1, 3, 5, and 10-fold cross-validations, respectively.

4.6. Instance retrieval

For the INSTRE dataset, saliency proposals quickly recall the objects that need to be retrieved because the dataset contains multiple target images. In this way, the results illustrate the higher values of mAP. On the other hand, the objects in most images are trademarks for the Flickr32 dataset. Our method outperforms all other methods. Compared to the mAP value on the INSTRE and Flickr32 datasets, the mAP obtained on the Flickr32 dataset is not as high because it could not effectively catch the key proposals in some images showing objects of smaller size. Fig. 10 shows the retrieved objects on the INSTRE dataset. The query objects are shown in orange, while retrieved objects are available in the green box. However, the red box shows the wrong object in the retrieved items. We do not have permission to show the retrieved objects from the Flickr32 dataset; therefore, we only show the mAP for the Flickr32 dataset.

5. Discussion

In this work, we proposed a novel salient object detection-based commodity items retrieval framework. Several key aspects are considered when designing the proposed framework regarding the choice of neural networks and methodology. We selected VGG16 in this retrieval framework because of several reasons. For example, VGG16 is known for extracting robust features from images. The network architecture, with its repeated blocks of convolutional layers followed by max-pooling, can capture a wide variety of visual features at different levels of abstraction. The VGG16 is pre-trained on a large dataset (ImageNet) consisting of millions of images with a thousand categories. This pre-training has allowed the VGG16 to learn a rich representation of visual features that can be beneficial when adapted to new tasks through transfer learning. The architecture of VGG16 is simple and uniform, making it easy to understand and modify. The consistent use of 3x3 convolutional filters throughout the network allows for a deeper architecture without a complicated design. Despite being surpassed by more recent architectures in terms of raw performance, VGG16 still performs very well on a variety of image-related tasks. Its performance is often good enough for many applications, especially when computational resources are limited. The network can be fine-tuned on specific tasks like salient object detection. Only the final layers must be trained from scratch, and the rest of the network can be slightly adjusted to better suit the specific data. VGG16 allows for end-to-end training, which can be advantageous when the salient object detection task requires learning complex patterns from low to high-level features.

It is important to note that while VGG16 has many advantages, it also has downsides, such as a large number of parameters, which can lead to high computational costs and the potential for overfitting on smaller datasets. However, when these drawbacks are not prohibitive,

Fig. 10. The qualitative performance of our proposed SODRet on INSTRE dataset. The images shown are a selection of query results, where the system has successfully identified and ranked relevant images that contain objects of interest similar to the query image.

Table 1
ime complexity of proposed work.

	Classification	Retrieval
Execution Time	600 ms	674 ms
Complexity Order	$O(n^2)$	$O(n^r)$

VGG16 can be a strong backbone for image retrieval systems based on salient object detection. We show the time complexity of VGG16 used in our retrieval task in Table 1.

Our proposed method also performed better for the INSTRE dataset across all validations as compared to the Flickr32 dataset due to its smaller number of classes and limited data. However, compared to other state-of-the-art methods, we achieved better performance across all validation scenarios.

6. Conclusion

In this study, we introduced a novel object retrieval framework that leverages a fusion-based approach for salient object detection—SODRet. Notably, SODRet operates independently of training data annotated with bounding boxes. Within our framework, we have developed a saliency detection network responsible for producing saliency maps. These maps are crucial as they enable the filtration of initial candidate proposals during the object retrieval process, thereby bolstering the recall rate of the intended targets by selectively narrowing down the pool of proposals.

Additionally, we have incorporated a unique feature mapping technique, referred to as the Channel-Weighted Generalized Mean Pooling (CWGMP) strategy, which is instrumental in maintaining the uniqueness of the convolutional features. This strategy effectively addresses the challenges associated with data burstiness and sporadic feature distribution across convolutional channels, enhancing the overall robustness and effectiveness of the retrieval system.

Through rigorous comparative experiments, our findings demonstrate that SODRet outperforms existing state-of-the-art methods, achieving a superior mAP, precision, recall, accuracy, and F1-score as well as ROC/AUC curves. We envision adapting SODRet for real-time online applications, particularly for consumer-oriented search tasks. The framework has the potential to be integrated with a personalized recommendation system that tailors search results to user preferences, offering a more curated and user-centric shopping experience.

CRediT authorship contribution statement

Muhammad Umair Hassan: Conceptualization, Methodology, Writing – original draft, Software, Writing – review & editing. **Xiuyang Zhao:** Supervision. **Raheem Sarwar:** Visualization, Investigation. **Naif R. Aljohani:** Data curation. **Ibrahim A. Hameed:** Software, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

We thank the anonymous reviewers for their valuable comments and suggestions in improving our manuscript. We also thank the Norwegian University of Science and Technology (NTNU), Norway, for supporting open access.

References

Babenko, A., & Lempitsky, V. (2015). Aggregating local deep features for image retrieval. In *Proceedings of the IEEE international conference on computer vision* (pp. 1269–1277).

Bhunia, A. K., Koley, S., Kumar, A., Sain, A., Chowdhury, P. N., Xiang, T., et al. (2023). Sketch2Saliency: Learning to detect salient objects from human drawings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2733–2743).

Cauteruccio, F., Corradini, E., Terracina, G., Ursino, D., & Virgili, L. (2022). Extraction and analysis of text patterns from NSFW adult content in reddit. *Data & Knowledge Engineering, 138*, Article 101979.

Chandrasekhar, V., Lin, J., Morère, O., Veillard, A., & Goh, H. (2015). Compact global descriptors for visual search. In *2015 Data compression conference* (pp. 333–342). IEEE.

Chaudhuri, U., Banerjee, B., Bhattacharya, A., & Datcu, M. (2020). CrossATNet-a novel cross-attention based framework for sketch-based image retrieval. *Image and Vision Computing, 104*, Article 104003.

Chen, Z., Kuang, Z., Wong, K.-Y. K., & Zhang, W. (2017). Aggregated deep feature from activation clusters for particular object retrieval. In *Proceedings of the on thematic workshops of ACM multimedia 2017* (pp. 44–51).

Cordonnier, J.-B., Mahendran, A., Dosovitskiy, A., Weissenborn, D., Uszkoreit, J., & Unterthiner, T. (2021). Differentiable patch selection for image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2351–2360).

Dost, S., Serafini, L., Rospocher, M., Ballan, L., & Sperduti, A. (2022). Aligning and linking entity mentions in image, text, and knowledge base. *Data & Knowledge Engineering*, *138*, Article 101975.

Fang, Z., Liu, J., Wang, Y., Li, Y., Hang, S., Tang, J., et al. (2016). Object-aware deep network for commodity image retrieval. In *Proceedings of the 2016 ACM on international conference on multimedia retrieval* (pp. 405–408).

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580–587).

Gudivada, V. N., & Raghavan, V. V. (1995). Content based image retrieval systems. *Computer*, *28*(9), 18–22.

Guo, J., & Lv, Z. (2022). Application of digital twins in multiple fields. *Multimedia Tools and Applications*, 1–27.

Hechavarria, A. A., & Shafiq, M. O. (2022). A modified attention mechanism powered by Bayesian network for user activity analysis and prediction. *Data & Knowledge Engineering*, Article 102034.

Hsu, T.-H., & Tang, J.-W. (2020). Development of hierarchical structure and analytical model of key factors for mobile app stickiness. *Journal of Innovation & Knowledge*, [ISSN: 2444-569X] *5*(1), 68–79. http://dx.doi.org/10.1016/j.jik.2019.01.006, URL: https://www.sciencedirect.com/science/article/pii/S2444569X19300204.

Jégou, H., Douze, M., Schmid, C., & Pérez, P. (2010). Aggregating local descriptors into a compact image representation. In *2010 IEEE computer society conference on computer vision and pattern recognition* (pp. 3304–3311). IEEE.

Ji, Y., Zhang, H., Zhang, Z., & Liu, M. (2021). CNN-based encoder-decoder networks for salient object detection: A comprehensive review and recent advances. *Information Sciences*, *546*, 835–857.

Kaur, J., & Singh, W. (2022). Tools, techniques, datasets and application areas for object detection in an image: a review. *Multimedia Tools and Applications*, 1–55.

Khanam, K. Z., Srivastava, G., & Mago, V. (2022). The homophily principle in social network analysis: A survey. *Multimedia Tools and Applications*, 1–44.

Kim, J., & Yoon, S.-E. (2018). Regional attention based deep feature for image retrieval. In *BMVC* (p. 209).

Kumar, R., et al. (2022). A hybrid feature extraction technique for content based medical image retrieval using segmentation and clustering techniques. *Multimedia Tools and Applications*, *81*(6), 8871–8904.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., et al. (2016). Ssd: Single shot multibox detector. In *European conference on computer vision* (pp. 21–37). Springer.

Liu, Y., Zhou, L., Bai, X., Huang, Y., Gu, L., Zhou, J., et al. (2021). Goal-oriented gaze estimation for zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3794–3803).

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision, vol. 2* (pp. 1150–1157). Ieee.

Lv, C., Zhou, X., Zhu, B., Liu, D., Zheng, B., Zhang, J., et al. (2023). SRI-Net: Similarity retrieval-based inference network for light field salient object detection. *Journal of Visual Communication and Image Representation*, *90*, Article 103721.

Mohedano, E., McGuinness, K., Giró-i Nieto, X., & O'Connor, N. E. (2018). Saliency weighted convolutional features for instance search. In *2018 International conference on content-based multimedia indexing* (pp. 1–6). IEEE.

Nasirtafreshi, I. (2022). Forecasting cryptocurrency prices using recurrent neural network and long short-term memory. *Data & Knowledge Engineering*, *139*, Article 102009.

Ozbay, F. A., & Alatas, B. (2021). Adaptive salp swarm optimization algorithms with inertia weights for novel fake news detection model in online social media. *Multimedia Tools and Applications*, *80*(26), 34333–34357.

Pirnay, L., & Burnay, C. (2022). How to build data-driven strategy maps? A methodological framework proposition. *Data & Knowledge Engineering*, *139*, Article 102019.

Rahman, M. A., & Wang, Y. (2016). Optimizing intersection-over-union in deep neural networks for image segmentation. In *International symposium on visual computing* (pp. 234–244). Springer.

Razavian, A. S., Sullivan, J., Carlsson, S., & Maki, A. (2016). Visual instance retrieval with deep convolutional networks. *ITE Transactions on Media Technology and Applications*, *4*(3), 251–258.

Revaud, J., Almazán, J., Rezende, R. S., & Souza, C. R. d. (2019). Learning with average precision: Training image retrieval with a listwise loss. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 5107–5116).

Romberg, S., Pueyo, L. G., Lienhart, R., & Van Zwol, R. (2011). Scalable logo recognition in real-world images. In *Proceedings of the 1st ACM international conference on multimedia retrieval* (pp. 1–8).

Salvador, A., Giró-i Nieto, X., Marqués, F., & Satoh, S. (2016). Faster r-CNN features for instance search. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 9–16).

Sang, J., Wu, Z., Guo, P., Hu, H., Xiang, H., Zhang, Q., et al. (2018). An improved YOLOv2 for vehicle detection. *Sensors*, *18*(12), 4272.

Sharif Razavian, A., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 806–813).

Sharif Razavian, A., Sullivan, J., Maki, A., & Carlsson, S. (2015). A baseline for visual instance retrieval with deep convolutional networks. In *International conference on learning representations*. ICLR.

Sivic, J., & Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *Computer vision, IEEE international conference on, vol. 3* (p. 1470). IEEE Computer Society.

Thomala, L. L. (2021). China: DAU of sina weibo 2021. *Statista*, URL: https://www.statista.com/.

Tolias, G., Avrithis, Y., & Jégou, H. (2016). Image search with selective match kernels: aggregation across single and multiple images. *International Journal of Computer Vision*, *116*(3), 247–261.

Tolias, G., & Jégou, H. (2014). Visual query expansion with or without geometry: refining local descriptors by feature aggregation. *Pattern Recognition*, *47*(10), 3466–3476.

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

Uijlings, J. R., Van De Sande, K. E., Gevers, T., & Smeulders, A. W. (2013). Selective search for object recognition. *International Journal of Computer Vision*, *104*(2), 154–171.

Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, *57*(2), 137–154.

Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., et al. (2018). Understanding convolution for semantic segmentation. In *2018 IEEE winter conference on applications of computer vision* (pp. 1451–1460). Ieee.

Wang, S., & Jiang, S. (2015). Instre: a new benchmark for instance-level object retrieval and recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *11*(3), 1–21.

Wang, W., Jiao, P., Liu, H., Ma, X., & Shang, Z. (2022). Two-stage content based image retrieval using sparse representation and feature fusion. *Multimedia Tools and Applications*, *81*(12), 16621–16644.

Wang, Z., Liu, X., Li, H., Shi, J., & Rao, Y. (2018). A saliency detection based unsupervised commodity object retrieval scheme. *IEEE Access*, *6*, 49902–49912.

Wang, L., Zhao, D., Wu, T., Fu, H., Wang, Z., Xiao, L., et al. (2020). Drosophila-inspired 3D moving object detection based on point clouds. *Information Sciences*, *534*, 154–171.

Wray, M., Larlus, D., Csurka, G., & Damen, D. (2019). Fine-grained action retrieval through multiple parts-of-speech embeddings. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 450–459).

Zhang, Y., Pan, P., Zheng, Y., Zhao, K., Zhang, Y., Ren, X., et al. (2018). Visual search at alibaba. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 993–1001).

Zhang, P., Wang, D., Lu, H., Wang, H., & Ruan, X. (2017). Amulet: Aggregating multi-level convolutional features for salient object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 202–211).

Zhang, Y. Y., Wang, H., Lv, X., & Zhang, P. (2021). Capturing the grouping and compactness of high-level semantic feature for saliency detection. *Neural Networks*, *142*, 351–362.

Zhang, G., Yang, Y., & Yang, G. (2022). Smart supply chain management in industry 4.0: the review, research agenda and strategies in North America. *Annals of Operations Research*, 1–43.

Zitnick, C. L., & Dollár, P. (2014). Edge boxes: Locating object proposals from edges. In *European conference on computer vision* (pp. 391–405). Springer.