Unsupervised Adversarial Instance-level Image Retrieval

C. Bai, H. Li, J. Zhang, L. Huang, Lu Zhang

Unsupervised Adversarial Instance-level Image Retrieval

Cong Bai, *Member, IEEE,* Hongkai Li, Jinglin Zhang*, Ling Huang and Lu Zhang

***Abstract*—With the wide use of visual sensors in the Internet of Things (IoT) in the past decades, huge amounts of images are captured in people's daily lives, which poses challenges to traditional deep-learning-based image retrieval frameworks. Most such frameworks need a large amount of annotated training data, which are expensive. Moreover, machines still lack human intelligence, as illustrated by the fact that they pay less attention to the interesting regions that humans generally focus on when searching for images. Hence, this paper proposes a novel unsupervised framework that focuses on the instance object in the image and integrates human intelligence into the deep-learning-based image retrieval. This framework is called adversarial instance-level image retrieval (AILIR). We incorporate adversarial training and an attention mechanism into this framework that considers human intelligence with artificial intelligence. The generator and discriminator are redesigned to guarantee that the generator retrieves similar images while the discriminator selects unmatched images and creates an adversarial reward for the generator. A minimax game is conducted by the adversarial reward retrieval mechanism until the discriminator is unable to judge whether the image sequence retrieved matches the query. Comparison and ablation experiments on four benchmark datasets prove that the proposed adversarial training framework indeed improves instance retrieval and outperforms the state-of-the-art methods focused on instance retrieval.**

***Index Terms*—Instance level image retrieval, Generative adversarial training, Human intelligence simulation, Unsupervised training**

I. Introduction

CONTENT-BASED image retrieval (CBIR) methods aim to find similar images in a dataset by extracting image visual information and have been studied since the 1970s [1]. The numbers of different kinds of images have vastly increased due to developments in the Internet of Things (IoT) [2], [3] and edge computing [4], [5], which bring more visual sensors into our daily lives. With this huge number of images, finding interesting images becomes more difficult. Thus, many challenges appear in CBIR. For example, given a query image containing an instance-level object, images containing similar instances are thought of as similar images that should be retrieved from the dataset. In this situation, the retrieval framework should focus on the instance-level objects during the retrieval process simulating the human intelligence, which is called instance-level image retrieval [6]. Scale-invariant feature transform (SIFT) [7], as a promising feature-extraction technology, and bag-of-visual-word (BoVW) [8], an advanced feature-representation framework, dominated instance-level image retrieval before the advent of deep-learning-based retrieval methods [9]. Generally speaking, in a deep-learning-based retrieval framework, compact feature vectors are extracted by a convolutional neural network (CNN), and the similarity between the query image and image dataset is measured by the approximate nearest neighbor (ANN). However, these frameworks need a large amount of labelled data for training, which takes significant time to produce. Furthermore, most of these frameworks lack an integration with human intelligence in the retrieval process, during which humans generally focus on the interesting regions. As a result, hybrid human-artificial intelligence has become a popular topic in deep-learning retrieval frameworks. Human intelligence can be used to optimize a framework so that it will pay attention to instance-level objects that are more meaningful to humans.

With the popularity of adversarial training, the generative adversarial network (GAN) [10] was proposed in 2014 to generate images similar to the input image. Many researchers then explored the possibilities of using the GAN in the applications related to images since it shows promising potential. There are three categories of image retrieval: class-level [11], [12], sketch-based [13], [14], and cross-modal retrieval [15], [16]. Most existing studies prefer to obtain more training data for image retrieval by employing the GAN. The GAN tends to generate synthetic images visually similar to the input image, which benefits the instance image retrieval task because the end goal is to retrieve similar images from a given dataset. While different approaches are adopted in the GAN and the retrieval task, both provide satisfactory results with despite having the same inputs. Therefore, it should be interesting and worthwhile to combine adversarial training and the retrieval procedure to redesign the GAN. The adversarial training can be treated as self-supervised or unsupervised training, which does not need supervised information in the training procedure. Furthermore, more weights should be added to the regions that are deemed interesting when humans see the images.

To address the above problems, a novel retrieval framework based on adversarial training is proposed in this paper, which is called adversarial instance-level image retrieval (AILIR). This framework takes advantage of hybrid human-artificial

This work is funded by National Natural Science Foundation of China under Grant No.61976192 and U1908210 and Zhejiang Provincial Natural Science Foundation of China under Grant No. LR21F020002. The source code of this work will be released on https://github.com/Zjut-MultimediaPlus.

Cong Bai, Hongkai Li and Ling Huang are with the College of Computer Science, Zhejiang University of Technology, Hangzhou 310023, China and are also with Key Laboratory of Visual Media Intelligent Processing Technology of Zhejiang Province, Hangzhou, China

Jinglin Zhang is with the School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China

Lu Zhang is with Univ Rennes, INSA Rennes, CNRS, IETR - UMR 6164, F-35000 Rennes, France

*Corresponding Author

Manuscript received xx; revised xx

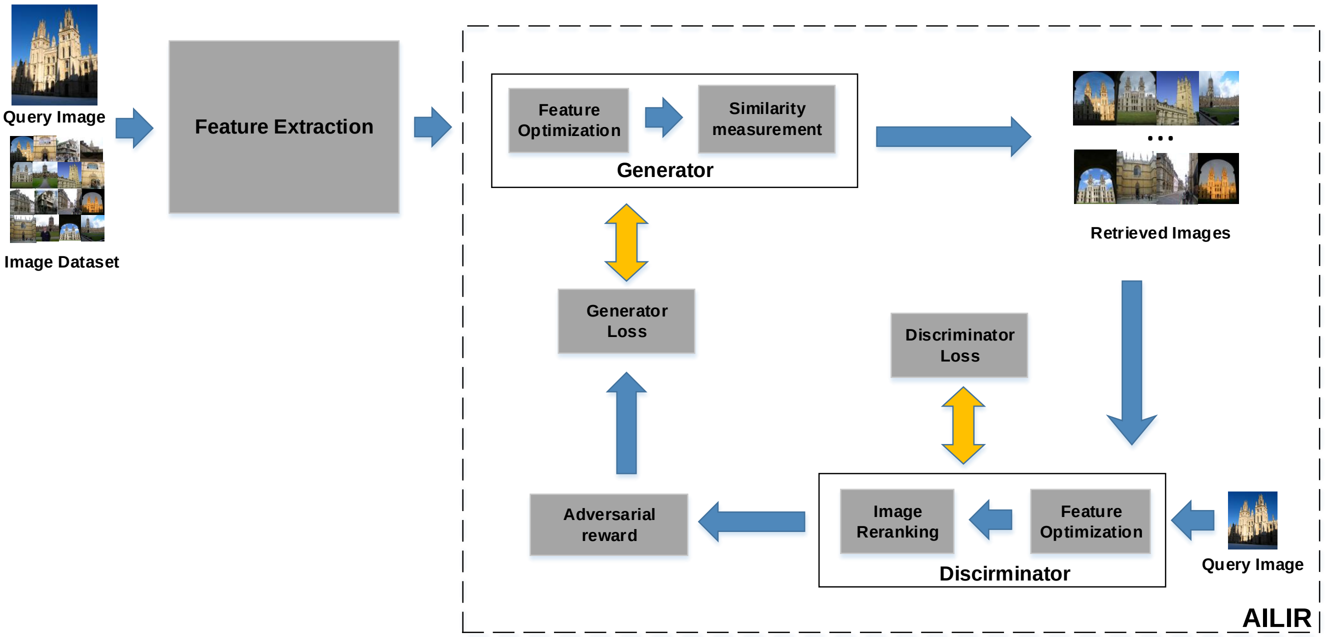

Fig. 1. Overview of the AILIR framework. Feature vectors of the query image and those of the dataset are used as the input of AILIR. Similar images are retrieved by the generator and the loss of generator is composed of image sequences retrieved, parameters of the generator network, and the adversarial reward. Images retrieved from the generator are judged by the discriminator for similarity to the input image. The discriminator loss is composed of image ranking and the parameters of the discriminator network. The judgement of similarity is output as the adversarial reward impacting on the generator loss to adjust the parameters of the generator.

intelligence by adopting adversarial training in the retrieval process and using human intelligence when humans try to find similar images. In the learning process, humans will first try their best to find similar images, and then they will learn some knowledge from the dissimilar images, which will help them to find more similar images during the next search. After many rounds of learning, they can finally find most of the similar images. Thus, in the proposed AILIR, given an image with a particular instance object as a query, images containing a similar instance object are output by the generator by optimizing features continuously to minimize the dissimilarities between the retrieved images and query image. The sequences of images are separated by the discriminator by maximizing the dissimilarity between the query image and retrieved images. An adversarial reward mechanism is proposed to link the parameter optimizations of the generator and discriminator. Finally, image retrieval is realized by the well-trained generator. In contrast to other deep-learning-based retrieval frameworks, AILIR training requires no human guidance or image labels. Hence, it can be considered an unsupervised training framework for instance-level image retrieval.

A preliminary version [17] of this article was presented at the 26th International Conference of Multimedia Modeling (MMM2020). Here we improve the training method for the framework and conduct comprehensive experiments in Section IV. We evaluate different inner structures of the generator/discriminator and conduct additional ablation experiments based on both hand-crafted features and deep dimensional features. Furthermore, we conduct a depth analysis and comparison with other state-of-the-art methods.

The contributions of our work are the following:

1) This paper proposes a novel end-to-end framework for instance-level image retrieval named AILIR. To the best of our knowledge, this is the first time that adversarial training has been adopted in a retrieval procedure for an instance-level image retrieval task. Furthermore, AILIR can be trained in an unsupervised way.

2) The generator and discriminator are redesigned with a 1×1 one-layer convolutional network for the retrieval task. An objective function and adversarial reward function are also proposed for adversarial retrieval training.

3) Comprehensive experiments evaluated using both on cross-validation and single-pass methods demonstrate that the incorporation of adversarial training in the retrieval process can significantly improve the retrieval accuracy without increasing time costs.

The remainder of this paper is organized as follows. Related work is introduced in Section II, and the details of AILIR are presented in section III. Section IV presents the experimental results, and conclusions are drawn in Section V.

II. Related Work

Deep learning has been applied in many fields such as image classification [18], image parsing [19], image retrieval [20], [21], [22], and cross-modal retrieval [23], [24], [25]. Furthermore, the GAN was proposed in 2014 to generate fake but similar images to the input image with the help of adversarial training. Since it is important for a deep-learning model to understand image data by learning its distribution, we briefly review the existing literature from two aspects in this section: deep-learning-based instance-level image retrieval and GAN-based image retrieval.

A. Deep-learning-based instance-level image retrieval

As introduced in Section I, the instance-level image retrieval task faces the challenge that similarity is measured at the object level. Many works have been dedicated to addressing this challenge. Initially, researchers have explored the instance-level retrieval task by extracting local descriptors to form global features to represent the instance object in the image. Those works include the following. Babenko et al. [26] aggregated local descriptors to generate compact global features and assemble features that were extracted from multiple scales. Kalantidis et al. [27] proposed a cross-dimensional weighted method that represents images in a set of deep networks. Meanwhile, Tuan et al. [28] used several masking methods to elect a representative subset of deep regional descriptors to make up global features. Next, several works were proposed that use information from important regions in images by extracting regional features to represent the instance object in the instance retrieval task. For example, class activation maps(CAM) [29] was proposed to acquire instance proposals after the most discriminative regions were obtained. Tolias et al. [30] improved global pooling approaches remarkably by region-of-interest-based pooling. Their approach was called regional Maximum Activation of Convolutions (R-MAC). This was the first attempt to divide an image into small blocks and to give a weight according to the background of the image. However, Kim et al. [31] found the problem that R-MAC considered many regions with meaningless backgrounds. To tackle this issue, Kim et al. proposed a simple yet effective regional attention network (R-mac + RA) that weighted an attentive score of a region considering the global context. Following these two milestone works, many researchers focused on activation maps [32], [33] and region attention weighting [34], [35] in instance-level retrieval.

For example, Xu et al. [36] utilized normalized feature maps as regional detectors to weigh and aggregate the convolutional features. This was the first attempt to combine aggregated representations and regional detectors. The regional detectors could highlight the discriminative parts of objects and effectively suppress background noise. Research on the class weighting network(CWN) [37] should also be highlighted. Based on semantic segmentation of image features, CWM encoded these features through a class weighting network and obtained the weight of each type of target by fine-tuning CNN classification. The next step was to recalculate region-wise weights between channels by using a spatial block. This study used image semantic segmentation technology to conduct successful image retrieval. Different from the aforementioned methods, some researchers investigated how to transform deep features into text so that they could be indexed with a standard text search engine [38]. However, most used deep-learning technology to extract different kinds of features, i.e., global, local, or regional.

B. GAN-based image retrieval

GAN has been widely used in image retrieval. There are two categories of GAN-based image retrieval methods: class-level image retrieval(CBIR) and sketch-based image retrieval(SBIR). Concerning the class-level image retrieval, HashGAN [39] can efficiently acquire image binary representations without pre-training. Moreover, a novel hashing loss function and a collaborative loss function were introduced to realize the similar random input and hash bit of the composite image. SSGAH [40] has three components: generative, discriminative, and deep hashing models. Generator and discriminator models are designed to learn triplet-wise information. As a result, binary codes in a hash model can obtain excellent semantic knowledge. BGAN [11] was proposed to convert images into binary codes in an unsupervised way. It also proposed a new sign-activation strategy and a loss function to solve the problem of how to equip the binary representation. Instead of using the GAN to generate binary representation in image retrieval, UAIR [41] was aimed at training the retrieval framework with unannotated information via adversarial learning. MindReader [42] was the first to use the GAN in the SBIR field. A cGAN [43] generated fake images using the given sketch image and output a learned encoder as a query sketch feature. ZS-SBIR [13] generated additional missing information for sketch images by using an adversarial auto-encoder and variational auto-encoder so as to retrieve more similar images. Meanwhile, the FHS-GAN [44] retrieved sketch images by generating freehand sketches in dual generative adversarial networks. CDRL [14] transformed sketches to images by employing the cycleGAN, aiming to learn more rich content relations between the sketch and image.

To the best of our knowledge, GAN-based instance-level image retrieval has never been conducted. The final step in the instance image retrieval task is to obtain images that have the same instance object as the input image. Thus, we attempt to integrate adversarial training and an instance retrieval stage with a re-designed generator and discriminator in order to achieve better retrieval performance.

III. Adversarial Training for Instance-level Image Retrieval

This paper presents an unsupervised instance-level image retrieval framework named AILIR. The generator and discriminator are redesigned by adopting adversarial training in the similarity measurement. The core details of the framework are presented in the following sections.

A. Framework Overview

The overall framework of the proposed AILIR is illustrated in Figure 1. As shown in the figure, the entire retrieval procedure is composed of two important stages. The first stage is feature extraction and the second is the AILIR framework, which is composed of a generator and discriminator. The generator retrieves images that contain similar instances as a given image. On the contrary, the discriminator judges whether the retrieved images have the specified instance. During training, these two components of AILIR play a minimax game via an adversarial reward. As a result, we can use the generator to complete instance-level image retrieval tasks.

B. Adversarial Reward Retrieval

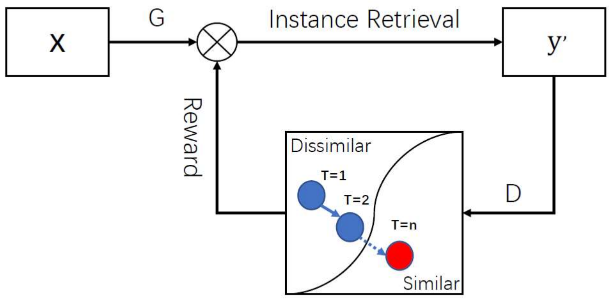

Fig. 2. Adversarial Reward Retrieval: At each training step, the generator retrieves the image y' while the discriminator discriminates whether the retrieved image is similar to the input image x and gives a reward to drive the generation for the next time step. The blue and red circles indicate the retrieved image. T represents the number of training epochs. After n training epochs, the image retrieved by the generator will become more similar.

Generative adversarial training offers a retrieval function that has been employed in classification [45], object recognition [46], and human pose estimation [47], [48]. The mechanism of adversarial reward retrieval is shown in Figure 2. In this figure, we can see that the images retrieved by the generator change from being dissimilar to the input image to being similar to the input image after adversarial training. This is a novel instance image retrieval method called adversarial reward retrieval. We now define our framework. The generator aims to iteratively improve its generations by using reward information produced by the discriminator. This is the first attempt to combine adversarial training and instance-level image retrieval. When the retrieved images are input to the discriminator, an adversarial reward, also called the relevance score, is established according to the re-ranked image sequences outputted by the discriminator. Furthermore, this adversarial reward optimizes the generator's loss function as a multiplied factor and updates the parameters of the generator so that it retrieves more similar images. The reward function is defined as follows:

$$reward = \delta(d_\phi(q, R)) \tag{1}$$

In the above equation, q represents the query image, and the images retrieved by the generator are denoted by R. δ is the sigmoid function, and d_ϕ is the cosine distance calculation function, which is defined as:

$$d_\phi(q|R) = \frac{\sum_{i=1}^{n} q_i R_i}{\sqrt{\sum_{i=1}^{n} q_i^2}\sqrt{\sum_{i=1}^{n} R_i^2}} \tag{2}$$

The adversarial reward is converted to the following after the sigmoid function is substituted as:

$$reward = \delta(d_\phi(q, R)) = \frac{exp(d_\phi(q, R))}{1 + exp(d_\phi(q, R))} \tag{3}$$

C. Novel Generator and Discriminator

To realize an instance-level image retrieval task, a novel framework was designed, and both the generator and discriminator use a 1×1 convolution kernel [49]. There are three reasons for this. First, the 1×1 convolution kernel is excellent for local patch feature extraction, which plays a crucial role in instance-level image retrieval. Second, image pixel information across channels can result in better integration, which can be seen as treating the local features of objects or scenes for retrieval in this task. Finally, in the training process, the number of channels in this network can be easily modified when we change a few network parameters.

Two components comprise the generator: the one-layer 1×1 convolutional network, which is used to optimize the input feature vectors, and the similarity measurement. With the new framework, it is easy to find the images from the gallery that have the same instance as the query. Thus, the generator model can be written as:

$$G_\theta(q, I) \tag{4}$$

where q represents the query image, I represents the photo gallery, and θ is the parameter of the generator. Before the distance calculation, the features of the query and dataset images are optimized by the network. Sorted by cosine similarity distance, the generator outputs the top k similar images.

Moreover, the goal of the discriminator is to determine whether the retrieved images have the same instance object. The network structure of the discriminator is similar to that of the generator. However, the input of the discriminator is the feature of input query and the top k retrieved images. As a result, the output is an adversarial reward. Features input into the discriminator are re-optimized by the generator and are different from those input into the generator. In the discriminator, a 1×1 convolutional network optimizes the feature vectors while an image re-ranking function adjusts the unmatched images. The discriminator model equation can be described as follows:

$$D_\phi(q, R) \tag{5}$$

where q denotes for the input query image, R represents the top similar retrieved images, and ϕ is the parameter of the discriminator. The discriminator outputs the image sequence that is re-ranked by the optimized feature vectors cosine distance.

D. Minimax Retrieval Game

In adversarial training, the generator and discriminator play a minimax game until the output of the discriminator is the

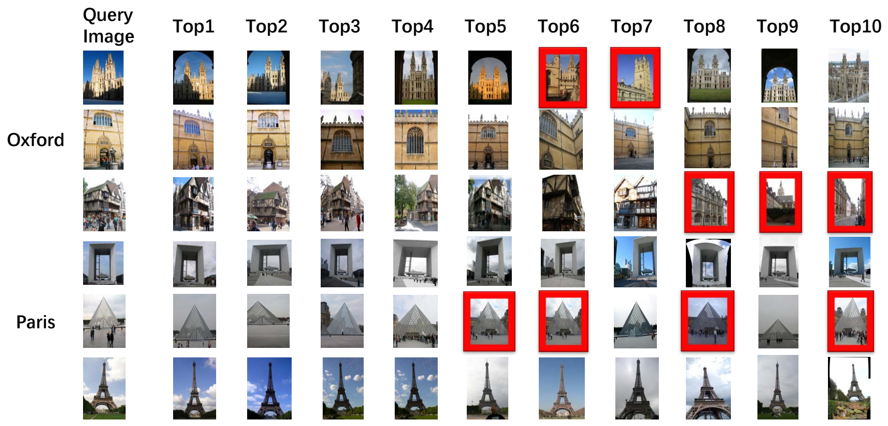

Fig. 3. Examples of images retrieved by the proposed AILIR on Oxford5K and Paris6K datasets. The top 10 retrieved images are shown. Inaccurate images are framed in red boxes.

same as that of the generator. The full objective function of AILIR is defined as follows:

$$\begin{aligned} f^{G,D} =& \min_{\theta}\max_{\phi} \sum_{n=1}^{N} \mathbb{E}_{R\sim G_{true(q_n,I)}}[logD_{\phi}(q_n,R)] \\ &+ \mathbb{E}_{R\sim G_{\theta(q_n,I)}}[log(1-D_{\phi}(q_n,R))] \end{aligned} \tag{6}$$

Where $G_{true(q_n,I))}$ expresses the perfect condition of the generator, which means that all of the images retrieved by the generator have the same instance as the query image. However, $G_{\theta(q_n,I)}$ indicates the current choice state for the generator. $\mathbb{E}$ represents the mathematical expectation of retrieved images R in the ground truth state or current state. In the optimization, θ must be minimized while ϕ must be maximized. This also means that the generator attempts to reduce the distance gap between the query image and the retrieved images; however, the discriminator tries to expand this distance gap constantly. The state is regarded as the ideal state when the calculated cosine distance is zero; this means that no additional training cost is required when adopting adversarial training.

In the full optimization function of AILIR, D_ϕ estimates the correlation between query image q and image dataset I or query image q and retrieved images R. Then, a relevance score is produced, which is same as the adversarial reward:

$$D_{\phi}(q_n,R) = \delta(d_{\phi}(q,R)) = \frac{exp(d_{\phi}(q,R))}{1+exp(d_{\phi}(q,R))} = reward \tag{7}$$

where q represents the query image feature and R represents the feature vectors of top similar images. q and R are inserted into the discriminator network for re-ranking. In the optimization of the discriminator for ϕ, Eq. 7 is substituted into Eq. 6, and then the discriminator objective function can be rewritten as:

$$\begin{aligned} f^{D} =& \underset{\phi}{argmax} \sum_{n=1}^{N} \mathbb{E}_{R\sim G_{true(q_n,I)}}[log\delta(d_{\phi}(q,R))] \\ &+ \mathbb{E}_{R\sim G_{\theta(q_n,I)}}[log(1-\delta(d_{\phi}(q,R))] \end{aligned} \tag{8}$$

The adversarial reward produced by the discriminator plays an important role in the training process of the generator. Only when the parameters θ are stable in the optimization can the generator be trained. If $D_\phi(q,R)$ is fixed, we can see that only the second part of Eq 6 must be optimized after Eq. 7 is substituted into Eq 6. Then, we can determine the objective function of the generator as follows:

$$\begin{aligned} f^{G} =& \underset{\theta}{argmin} \sum_{n=1}^{N} \mathbb{E}_{R\sim G_{true(q_n,I)}}[log\delta(d_{\phi}(q,R))] \\ &+ \mathbb{E}_{R\sim G_{\theta(q_n,I)}}[log(1-\delta(d_{\phi}(q,R))] \\ =& \underset{\phi}{argmax} \sum_{n=1}^{N} \mathbb{E}_{R\sim G_{\theta(q_n,I)}}[log(1+\delta(d_{\phi}(q,R))] \end{aligned} \tag{9}$$

As shown above, we know that the $reward = \delta(d_\phi(q,R)$ is treated as a penalty term in the generator objective function. The discriminator objective function is continuous, and the stochastic gradient descent method is used for parameter optimization of the discriminator. On the contrary, the generator objective function is a discrete function, and hence we adopt the advanced gradient algorithm (Adam) to optimize the generator.

E. Training AILIR

We summarize the adversarial training between the two components in the retrieval process in Algorithm 1. The generator and discriminator should be initialized before training starts. Then, the generator and discriminator are trained alternatively via Eqs. 8 and Eqs 9 during the adversarial training stage.

IV. EXPERIMENTS

We conducted experiments to evaluate the performance of the AILIR framework. These experiments were designed from three aspects and tested on four widely used benchmark datasets. First, we evaluated the different inner structures of the generator/discriminator, including the numbers of layers

Algorithm 1: Adversarial training for instance image retrieval

```
Input: The feature vectors of the image database and that of the query image;
Initialize the weights of the generator $G(s')$ and discriminator $D(s)$;
repeat
    for epoch do
        The generator $G(s')$ retrieves the similar images from the database to train the discriminator $D(s)$;
        for d-epochs do
            Calculate the distance between the query image and the sequence of images retrieved by the generator and obtain the reward;
            Optimize the parameters of $D(s)$ by the stochastic gradient descent;
        end
        for g-epochs do
            Calculate the distance between the query image and the gallery images;
            Optimize the parameters of $G(s')$ by Adam algorithm;
        end
    end
until convergence;
```

and number of channels. Moreover, an ablation experiment for adversarial training was conducted. Finally, AILIR was compared with other state-of-the-art methods. The details of experiments are shown in the following.

A. Datasets and Evaluation Criteria

The Oxford5K [50] dataset contains 5,063 images of 11 Oxford landmarks, obtained from Flickr. This dataset defines five queries for each landmark through hand-drawn bounding boxes, resulting in a total of 55 regions of interest (ROIs). Each image is assigned one of four labels, good, ok, junk, or bad. The good and ok labels indicate that the ROIs of the query are well-matched, while the bad label means they are not matched.

The Paris6K [51] dataset contains 6,412 images of 11 Paris landmarks. This database also has five queries per landmark, and thus 55 queries with bounding boxes. Moreover, the labels are the same as those in the Oxford5K dataset.

The Flickr100K [51] dataset is formed by 99,782 images marked with 145 famous landmark labels. To obtain the expandability performance in retrieval, this dataset is usually added to Oxford5k and Paris6K to compose Oxford105K or Paris106K, respectively.

As others have done, we measured the retrieval result by the mean average precision (mAP), which represents the average percentage of the same-class images in all retrieved images after evaluating all queries. The K average precision formula is recommended as follows:

$$precision@K = \frac{\sum_{i=1}^{k} R_e(i)}{K} \tag{10}$$

where $R_e(i) \in [0, 1]$ indicates the ground truth and determines the relevance between the query image and the i-th ranked image. 0 means that the image does not have the same instance-level object as the query image and 1 means that it does. The mAP formula is defined as follows:

$$mAP(Q) = \frac{1}{|Q|} \sum_{i=0}^{Q} \frac{1}{m} \sum_{k=1}^{m} precision \tag{11}$$

where Q is the set of query images and m is the number of relevant images in the dataset.

B. Experimental Details

1) Network Settings: Because of the simplicity and low computing burden of the 1×1 one-layer convolution network, it is used in both components of the AILIR framework. Random normal initialization is used to initialize the network parameters. The numbers of layers and the number of channels are varied, as discussed in Sections IV-D1 and IV-D2.

2) Framework Input: Any type of feature can be inputted into framework, such as the hand-crafted features, CNN-based deep dimensional features, or regional features. In the ablation experiments, we used deep features and hand-crafted features for the input. Since regional features [31] were well recognized in instance-level retrieval, they are used to evaluate the performance of AILIR in comparison with other start-of-the-art methods. Before starting to train the AILIR framework, no label information should be annotated for images. Thus, we claim that we trained the AILIR in an unsupervised way.

3) Framework Output: Finally, the outputs of AILIR are the retrieved images that have the same instance as the query image.

C. Framework Training

After the features are input to the AILIR, the network parameters are randomly initialized. We set the learning rate of the generator as 8×10^{-3} and that of the discriminator as 3×10^{-3}. We decayed them by 0.1 for every 10 epochs. All methods were trained for 30 epochs, and we evaluated the metrics on the validation set to select the best model for all methods after every epoch.

To simplify the training process and examine the robustness of AILIR, we evaluated the performance in a pre-trained single-pass way [6]. That is to say, we used the Landmark dataset [52] to train the AILIR, and retrieval was evaluated on the Paris or Oxford dataset. The experimental results reported in Tables I- III were obtained in this way. These results are also indicated by AILIR (Single) in Table IV.

Furthermore, to make a fair comparison with other methods, we used cross-validation evaluation, which is the same as R-MAC [30], CroW [27] and SDCF [28]. That is to say, when the retrieval performance of AILIR was evaluated on Oxford5K and Oxford105K, the Paris6K dataset was used as a training dataset, and vice versa. The results obtained in this way are marked as AILIR (Cross) in Table IV. Generally speaking, the performance obtained in this way is better than that obtained in the single-pass way, as the objects in Oxford5K are similar to the objects in Paris6K.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3065578, IEEE Transactions on Multimedia

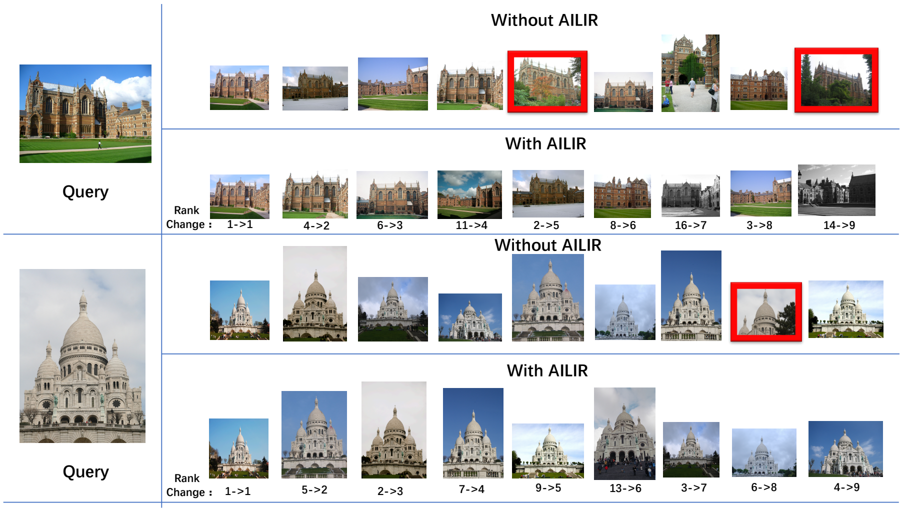

Fig. 4. Retrieval results with and without AILIR for the same query image. Ranking changes are shown as “A->B”. A indicates the ranking number without AILIR, and B that with AILIR. The retrieved images that do not have same label as the query are marked by red boxes. ”Without AILIR” means that method ”R-MAC + RA” was used to retrieve the image, which is served as one of the baseline methods. ”With AILIR” means that method ”R-MAC + RA + AILIR” was used.

TABLE I
DIFFERENT NUMBERS OF 1 × 1 CONVOLUTIONAL LAYERS

Number of layers	mAP(%)	
	Oxford5k	Paris6k
One	**79.25**	**86.93**
Two	74.98	86.13
Three	73.83	85.32

TABLE II
DIFFERENT NUMBERS OF 1 × 1 CONVOLUTIONAL CHANNELS

Number of channels	mAP(%)		time(s)	
	Oxford5k	Paris6k	Oxford5k	Paris6K
2048	**79.25**	**86.93**	0.145	0.163
1024	75.9	86.13	0.109	0.127
512	71.04	84.26	0.109	0.127
256	63.86	82.48	0.091	0.109
128	52.7	74.3	0.091	0.091

D. Experimental Results

Several examples of retrieved images are shown in Figure 3. Inaccurate images are marked by red boxes. We can see that some inaccurate images are somewhat similar to the query image, e.g., those in the fifth row in Figure 3. This is mainly because the inaccurate image contains a similar instance but does not share the same label as the query image.

1) Different number of layers in the convolutional network: The convolutional networks of the generator and discriminator can be composed of different numbers of layers. We increased the number of layers from 1 to 3 to observe the changes in image retrieval performance. Table I shows the results of these experiments conducted on the Oxford5K and Paris6K datasets.

From Table I, we can see that the number of layers impacts the retrieval performance. The reason for this is that a different number of parameters will be learned during the training with each different number of layers. Such differences will lead to different learning abilities. We can see that a one-layer convolutional network achieves the best performance for both datasets. With a greater number of layers, more parameters must be trained, which means that training may sometimes fall into a local optimum sometimes. However, a one-layer 1×1 convolution network did not increase the retrieval burden obviously, so it was used in the subsequent experiments.

2) Different number of channels in the convolutional network: The dimension of an output feature vector can be reduced or increased by the 1×1 convolutional layer. To clarify how the retrieval performance is influenced by changes in the feather vector dimension, we varied the dimension from 128 to 2048. The features inputted in this experiment were the 2048-dimensional local features extracted in the regional proposal network. Table II shows the experimental results.

It can conclude from Table II that a decrease in the number of dimensions will slightly reduce the retrieval performance. However, when the number of dimensions is significantly smaller, e.g., 128 or 256, mAP greatly decreases. Therefore, we could say that AILIR remains robust when the number of dimensions is within a reasonable range. To obtain better experimental results, 2048-dimensional feature vectors were used in the adversarial training in subsequent experiments.

Concerning the retrieval time, also shown in Table II, the feature vector with a large dimension will be slightly longer than that with a small dimension. To obtain better experimental results, we used 2048-dimensional feature vectors in subsequent experiments.

3) AILIR ablation experiments: One of our main contributions is to incorporate adversarial training in the retrieval process. We conducted ablation experiments to determine whether adding adversarial training improves performance. To realize this, the traditional hand-crafted features such as GIST [53] and the CNN-based features such as AlexNet [54] were both used as the ablation baseline. Furthermore, two milestone methods for instance-level image retrieval were also used: R-MAC [30] and its improved version with regional attention [31] (R-MAC+RA). To make the comparison, we

TABLE III
AILIR ABLATION EXPERIMENT RESULTS

Methods	mAP(%)		Time(s)	
	Oxford5k	Paris6k	Oxford5k	Paris6K
GIST	29.39	20.41	0.127	0.164
GIST + AILIR	**32.57**	**21.48**	0.145	0.181
AlexNet	44.26	67.21	0.109	0.127
AlexNet + AILIR	**46.83**	**69.41**	0.145	0.164
R-MAC	70.01	85.4	0.091	0.109
R-MAC + AILIR	**74.16**	**85.9**	0.109	0.127
R-MAC + RA	76.8	**87.5**	0.091	0.109
R-MAC + RA + AILIR	**79.25**	86.93	0.145	0.164

obtained four types of features, i.e., GIST, AlexNet, R-MAC and R-MAC+RA from their proposed network, and they were regarded as the AILIR input.

It can be seen from the results in Table III that adopting the adversarial training increases the time cost by approximately 0.02-0.04 s. Therefore, we affirm that adding adversarial training indeed improves retrieval performance but with a slightly extra time cost.

Figure 4 visualizes the changes in the retrieved image sequence after applying AILIR. Although the image sequence order changes, the similar images are still in the top ranking. However, retrieved images with different labels disappear when AILIR is applied.

4) Comparison with state-of-the-art methods: We used several retrieval methods for comparison that focus on instance retrieval, including SPoC [26], GatedSQU [32], ReSW[34], AdCoW[35], CWN[37], BoDVW [33], SDCF [28], Crow [27], R-MAC [30], R-MAC + RA [31] and PWA [36]. The experiments were conducted on the publicly available datasets Oxford5k, Paris6k, Oxford105k and Paris106k. For the sake of fairness, the dimensions of image features used in the AILIR were set to be the same as its competitors, either 512 or 2048 dimensions. Table IV compares the comparisons of the performance on four datasets.

If a 512-dimensional feature vector is used, the proposed AILIR achieves the best results for Paris6k and Paris106k among the compared methods; meanwhile, its performance for Oxford5K and Oxford105k is slightly lower than the best performance. If a 2048-dimensional feature vector is used, we could see that the proposed AILIR (with both single-pass and cross-validation training) acquires the best overall retrieval performance for Oxford5K, Oxford105K and Pairs106K, with a 3% to 6% improvement over the second-best state-of-the-art method. AILIR performance is second-best, with a slight gap in retrieval with respect to Paris6K. However, it obviously outperforms all other methods. We found that the proportion of instance objects in Paris6K query images is smaller than that in Oxford5K query images, which leads to the objects in Paris6K being more obvious. This is why the overall retrieval performance for Paris6k is better than that for Oxford5K. However, from the last three rows in Table IV, we can see that the proposed AILIR still presents an obvious improvement for Oxford5K, which proves the advancement made with AILIR. The difference in performance between training in a single-pass way and training in a cross-validation way is slight, which illustrates the robustness of AILIR in training methods. In conclusion, adopting adversarial training in the retrieval process could improve performance on in instance-level image retrieval tasks.

V. Conclusions

We propose an unsupervised framework called AILIR for instance-level image retrieval to adopt adversarial training in retrieval procedure rather than data augmentation. To realize this, adversarial reward function, generator and discriminator are redesigned for retrieval purposes. AILIR is trained in an unsupervised way with no annotated information available during training. In the experimental verification stage, comprehensive experiments are conducted to assess different inner structures of the generator/discriminator, the ablation experiments of adversarial training and the retrieval performance comparison with other state-of-the-art methods. It is clear that adversarial training indeed works for instance-level image retrieval with slightly higher time consumption; moreover, AILIR could achieve better performance than existing methods that focus on instance-level retrieval. However, a lightweight and pruned model is expected to be useful in the field of mobile image retrieval, and thus we aim to compress AILIR for mobile applications in the future.

References

[1] Y. Rui, T. S. Huang, and S.-F. Chang, “Image Retrieval: Current Techniques, Promising Directions, and Open Issues,” *Journal of Visual Communication and Image Representation*, vol. 10, no. 1, pp. 39–62, mar 1999.

[2] F. Shi, H. Ning, W. Huangfu, F. Zhang, D. Wei, T. Hong, and M. Daneshmand, “Recent progress on the convergence of the internet of things and artificial intelligence,” *IEEE Network*, vol. 34, no. 5, pp. 8–15, 2020.

[3] J. Zhang, M. Dai, and Z. Su, “Task allocation with unmanned surface vehicles in smart ocean iot,” *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 9702–9713, 2020.

[4] H. Ning, X. Liu, X. Ye, J. He, W. Zhang, and M. Daneshmand, “Edge computing-based id and nid combined identification and resolution scheme in iot,” *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 6811–6821, 2019.

[5] P. Hu, W. Chen, C. He, Y. Li, and H. Ning, “Software-defined edge computing (sdec): Principle, open iot system architecture, applications and challenges,” *IEEE Internet of Things Journal*, 2019.

[6] L. Zheng, Y. Yang, and Q. Tian, “Sift meets cnn: A decade survey of instance retrieval,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 5, pp. 1224–1244, 2017.

[7] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, nov 2004.

[8] Sivic and Zisserman, “Video Google: a text retrieval approach to object matching in videos,” in *Proceedings Ninth IEEE International Conference on Computer Vision*, vol. 2. IEEE, 2003, pp. 1470–1477.

[9] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li, “Deep learning for content-based image retrieval: A comprehensive study,” in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 157–166.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[11] J. Song, T. He, L. Gao, X. Xu, A. Hanjalic, and H. T. Shen, “Binary Generative Adversarial Networks for Image Retrieval,” in *AAAI Conference on Artificial Intelligence*, 2018, pp. 394–401.

[12] L. Huang, C. Bai, Y. Lu, S. Chen, and Q. Tian, “Adversarial learning for content-based image retrieval,” in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2019.

TABLE IV
COMPARISON WITH STATE-OF-THE-ART METHODS

Methods	Dim.	mAP(%)			
		Oxford5k	Paris6k	Oxford105k	Paris106K
SPoC[26]	512	66.9	76.5	-	-
GatedSQU[32]	512	69.4	81.3	63.9	73.4
R-MAC[30]	512	67.0	83.0	61.6	75.7
CAM[29]	512	71.2	80.5	67.2	73.3
ReSW[34]	512	72.6	82.4	67.5	73.0
AdCoW[35]	512	72.8	83.0	68.1	76.3
CWN[37]	512	**73.48**	84.98	**70.18**	78.89
AILIR(Single)	512	71.04	84.26	69.72	83.88
AILIR(Cross)	512	70.7	**85.2**	68.7	**84.7**
BoDVW[33]	Bow-10k	77.6	85.0	74.9	79.4
SDCF[28]	2048	69.1	81.7	65.4	74.3
Crow[27]	2048	68.7	82.8	62.7	75.1
R-MAC[30]	2048	70.1	85.4	66.9	80.8
PWA[36]	2048	77.6	84.5	71.1	78.2
R-MAC + RA[31]	2048	76.8	**87.5**	73.6	82.5
AILIR (Single)	2048	79.25	86.83	78.4	85.89
AILIR (Cross)	2048	**79.4**	87.38	**79.2**	**86.6**

[13] S. K. Y. B, S. K. Reddy, and A. Mishra, “A Zero-Shot Framework for Sketch Based Image Retrieval,” in *ECCV 2018*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11219. Springer International Publishing, 2018, pp. 316–333.

[14] C. Bai, J. Chen, Q. Ma, P. Hao, and S. Chen, “Cross-domain representation learning by domain-migration generative adversarial network for sketch based image retrieval,” *Journal of Visual Communication and Image Representation*, p. 102835, 2020.

[15] X. Huang, Y. Peng, and M. Yuan, “Mhtn: Modal-adversarial hybrid transfer network for cross-modal retrieval,” *IEEE Transactions on Cybernetics*, pp. 1–13, 2018.

[16] C. Bai, C. Zeng, Q. Ma, J. Zhang, and S. Chen, “Deep adversarial discrete hashing for cross-modal retrieval,” in *Proceedings of the 2020 International Conference on Multimedia Retrieval*, ser. ICMR2020. New York, NY, USA: Association for Computing Machinery, 2020, pp. 525–531.

[17] H. Li, C. Bai, L. Huang, Y. Jiang, and S. Chen, “Instance image retrieval with generative adversarial training,” in *International Conference on Multimedia Modeling*. Springer, 2020, pp. 381–392.

[18] C. Zhang, J. Cheng, and Q. Tian, “Unsupervised and semi-supervised image classification with weak semantic consistency,” *IEEE Transactions on Multimedia*, vol. 21, no. 10, pp. 2482–2491, 2019.

[19] R. Zhao, Y. Xue, J. Cai, and Z. Gao, “Parsing human image by fusing semantic and spatial features: A deep learning approach,” *Information Processing and Management*, vol. 57, no. 6, p. 102306, 2020.

[20] C. Bai, L. Huang, X. Pan, J. Zheng, and S. Chen, “Optimization of deep convolutional neural network for large scale image retrieval,” *Neurocomputing*, vol. 303, pp. 60 – 67, 2018.

[21] R. Hong, L. Li, J. Cai, D. Tao, M. Wang, and Q. Tian, “Coherent semantic-visual indexing for large-scale image retrieval in the cloud,” *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4128–4138, 2017.

[22] F. Yang, Y. Wu, Z. Wang, X. Li, S. Sakti, and S. Nakamura, “Instance-level heterogeneous domain adaptation for limited-labeled sketch-to-photo retrieval,” *IEEE Transactions on Multimedia*, 2020.

[23] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, “Generalized semi-supervised and structured subspace learning for cross-modal retrieval,” *IEEE TRANSACTIONS ON MULTIMEDIA*, vol. 20, no. 1, pp. 128–141, 2018.

[24] W. Wang, J. Gao, X. Yang, and C. Xu, “Learning coarse-to-fine graph neural networks for video-text retrieval,” *IEEE Transactions on Multimedia*, pp. 1–1, 2020.

[25] J. Chen, L. Zhang, C. Bai, and K. Kpalma, “Review of recent deep learning based methods for image-text retrieval,” in *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2020, pp. 167–172.

[26] A. Babenko and V. Lempitsky, “Aggregating local deep features for image retrieval,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1269–1277.

[27] Y. Kalantidis, C. Mellina, and S. Osindero, “Cross-dimensional weighting for aggregated deep convolutional features,” in *European conference on computer vision*. Springer, 2016, pp. 685–701.

[28] T. Hoang, T.-T. Do, D.-K. Le Tan, and N.-M. Cheung, “Selective deep convolutional features for image retrieval,” in *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 2017, pp. 1600–1608.

[29] A. Jimenez, J. M. Alvarez, and X. Giro-i Nieto, “Class-weighted convolutional features for visual instance search,” *arXiv preprint arXiv:1707.02581*, 2017.

[30] G. Tolias, R. Sicre, and H. Jégou, “Particular object retrieval with integral max-pooling of CNN activations,” in *International Conference on Learning Representations (ICRL)*, San Juan, Puerto Rico, 2016, pp. 1–12.

[31] J. Kim and S.-E. Yoon, “Regional attention based deep feature for image retrieval,” in *British Machine Vision Conference (BMVC)*. BMVA, 2018, pp. 1–13.

[32] Z. Chen, J. Lin, V. Chandrasekhar, and L.-Y. Duan, “Gated square-root pooling for image instance retrieval,” in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 1982–1986.

[33] Y. Lv, W. Zhou, Q. Tian, and H. Li, “Scalable bag of selected deep features for visual instance retrieval,” in *International Conference on Multimedia Modeling*. Springer, 2018, pp. 239–251.

[34] S. Pang, J. Zhu, J. Wang, V. Ordonez, and J. Xue, “Building discriminative cnn image representations for object retrieval using the replicator equation,” *Pattern Recognition*, vol. 83, pp. 150–160, 2018.

[35] J. Zhu, J. Wang, S. Pang, W. Guan, Z. Li, Y. Li, and X. Qian, “Co-weighting semantic convolutional features for object retrieval,” *Journal of Visual Communication and Image Representation*, 2019.

[36] J. Xu, C. Shi, C. Qi, C. Wang, and B. Xiao, “Unsupervised part-based weighting aggregation of deep convolutional features for image retrieval,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018, pp. 7436–7443.

[37] X. Li, K. Jin, and R. Long, “End-to-end semantic-aware object retrieval based on region-wise attention,” *Neurocomputing*, vol. 359, pp. 219 – 226, 2019.

[38] G. Amato, F. Carrara, F. Falchi, C. Gennaro, and L. Vadicamo, “Large-scale instance-level image retrieval,” *Information Processing and Management*, p. 102100, 2019.

[39] K. Ghasedi Dizaji, F. Zheng, N. Sadoughi, Y. Yang, C. Deng, and H. Huang, “Unsupervised deep generative adversarial hashing network,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3664–3673.

[40] G. Wang, Q. Hu, J. Cheng, and Z. Hou, “Semi-supervised generative adversarial hashing for image retrieval,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 469–485.

[41] L. Huang, C. Bai, Y. Lu, S. Chen, and Q. Tian, “Adversarial learning for content-based image retrieval,” in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, March 2019, pp. 97–102.

[42] L. Guo, J. Liu, Y. Wang, Z. Luo, W. Wen, and H. Lu, “Sketch-based image retrieval using generative adversarial networks,” in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1267–1268.

[43] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” *arXiv preprint arXiv:1411.1784*, 2014.

[44] X. Zhang, X. Li, X. Li, and M. Shen, "Better freehand sketch synthesis for sketch-based image retrieval: Beyond image edges," *Neurocomputing*, vol. 322, pp. 38–46, 2018.

[45] A. R. Zamir, T.-L. Wu, L. Sun, W. B. Shen, B. E. Shi, J. Malik, and S. Savarese, "Feedback networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1308–1317.

[46] M. Liang and X. Hu, "Recurrent convolutional neural network for object recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3367–3375.

[47] V. Belagiannis and A. Zisserman, "Recurrent human pose estimation," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017).* IEEE, 2017, pp. 468–475.

[48] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4733–4742.

[49] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[50] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[51] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, June 2008, pp. 1–8.

[52] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *Lecture Notes in Computer Science*, vol. 8689 LNCS, 2014, pp. 584–599.

[53] A. Friedman, "Framing pictures: The role of knowledge in automatized encoding and memory for gist." *Journal of experimental psychology: General*, vol. 108, no. 3, p. 316, 1979.

[54] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.