IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 November 2022, accepted 5 January 2023, date of publication 2 February 2023, date of current version 2 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3241858

RESEARCH ARTICLE

Image Retrieval Using Convolutional Autoencoder, InfoGAN, and Vision Transformer Unsupervised Models

EMAN S. SABRY[1], SALAH S. ELAGOOZ[1], FATHI E. ABD EL-SAMIE[2], WALID EL-SHAFAI[2,3], NIRMEEN A. EL-BAHNASAWY[4], GHADA M. EL-BANBY[5], ABEER D. ALGARNI[6], NAGLAA F. SOLIMAN[6], AND RABIE A. RAMADAN[7], (Member, IEEE)

[1]Department of Communications and Computers Engineering, Higher Institute of Engineering, El-Shorouk Academy, El-Shorouk 11837, Egypt
[2]Department of Electronics and Electrical Communications Engineering, Faculty of Electronic Engineering, Menoufia University, Menouf 32952, Egypt
[3]Security Engineering Laboratory, Computer Science Department, Prince Sultan University, Riyadh 11586, Saudi Arabia
[4]Computer Science and Engineering Department, Faculty of Electronic Engineering, Menoufia University, Menouf 32952, Egypt
[5]Department of Industrial Electronics and Control Engineering, Faculty of Electronic Engineering, Menoufia University, Menouf 32952, Egypt
[6]Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia
[7]Computer Engineering Department, College of Engineering, Cairo University, Giza 12613, Egypt

Corresponding author: Walid El-Shafai (eng.waled.elshafai@gmail.com)

This work was supported by Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia, through the Princess Nourah bint Abdulrahman University Researchers Supporting Project, under Project PNURSP2023R66.

ABSTRACT Query by Image Content (QBIC), subsequently known as Content-Based Image Retrieval (CBIR), offers an advantageous solution in a variety of applications, including medical, meteorological, search by image, and other applications. Such CBIR systems primarily use similarity matching algorithms to compare image content to get matched images from datasets. They essentially measure the spatial distance between extracted visual features from a query image and its similar versions in the dataset. One of the most challenging query retrieval problems is Facial Sketched-Real Image Retrieval (FSRIR), which is based on content similarity matching. These facial retrieval systems are employed in a variety of contexts, including criminal justice. The difficulties of retrieving such sorts come from the composition of the human face and its distinctive parts. In addition, the comparison between these types of images is made within two different domains. Besides, to our knowledge, there is a few large-scale facial datasets that can be used to assess the performance of the retrieval systems. The success of the retrieval process is governed by the method used to estimate similarity and the efficient representation of compared images. However, by effectively representing visual features, the main challenge-posing component of such systems might be resolved. Hence, this paper has several contributions that fill the research gap in content-based similarity matching and retrieval. The first contribution is extending the Chinese University Face Sketch (CUFS) dataset by including augmented images, introducing to the community a novel dataset named Extended Sketched-Real Image Retrieval (ESRIR). The CUFS dataset has been extended from 100 images to include 53,000 facial sketches and 53,000 real facial images. The paper second contribution is presenting three new systems for sketched-real image retrieval based on convolutional autoencoder, InfoGAN, and Vision Transformer (ViT) unsupervised models for large datasets. Furthermore, to meet the subjective demands of the users due to the prevalence of multiple query formats, the third contribution of the paper is to train and assess the performance of the proposed models on two additional facial datasets of different image types. Recently, the majority of people have preferred searching for brand logo images, but it may be tricky to separate certain brand logo features their alternatives and even from other features in an image. Thus, the fourth contribution is to compare logo

The associate editor coordinating the review of this manuscript and approving it for publication was Abdel-Hamid Soliman.

image retrieval performance based on visual features derived from each of the three suggested retrieval systems. The paper also presents cloud-based energy and computational complexity saving approaches on large-scale datasets. Due to the ubiquity of touchscreen devices, users often make drawings based on their fantasies for certain object image searches. Thus, the proposed models are tested and assessed on a tough dataset of doodle-scratched human artworks. They are also studied on a multi-category dataset to cover practically all possible image types and situations. The results are compared with those of the most recent algorithms found in the literature. The results show that the proposed systems outperform the recent counterparts.

INDEX TERMS Feature extraction, InfoGAN, sketched-real image retrieval, object matching, spatial distance measurement, vision transformer.

I. INTRODUCTION

A. OVERVIEW

Generally, feature extraction, indexing, and similarity measurement are the primary building blocks for any image retrieval system. In such types of systems, the feature extraction block serves as the key engine for the rest of the blocks [3]. The capacity of such technologies to generate appropriate results is constrained by the refinement of the image content [4]. The need to manage vast amounts of features representing details in images raises a slew of enforcement hurdles. Furthermore, the availability of a diversity of quality and query forms [5] leads to the appearance of numerous data choices, making it harder to match the user's intent with high retrieval speed and accuracy. In addition, the large size of the image datasets decreases the system computational efficiency, accuracy, and memory saving. Indeed, the redundancy level in images significantly impacts the image storage size and retrieved features using extraction methods. As a result, the more discriminative the image representation with suitable feature descriptor dimensionality, the better the accuracy of image matching is.

Therefore, the applicable feature extraction algorithm, which allows such image representation, becomes more successful. Image features might be represented globally or locally. The complete image contour is represented by global features, which may be utilized to detect duplication in a large-scale image dataset [6]. On the other hand, a local feature representation is a pattern or unique structure detected in an image, such as a point, an edge, or a tiny image patch. Such approaches portray the contents by focusing on a few key regions within the same image that might differ from region to region. These regions are unaffected by changes in perspective and lighting. Content-Based Image Retrieval (CBIR) and other applications rely largely on such local fine details or features [7], [8]. Consequently, the most significant issue to consider is how such features are extracted.

Convolutional Neural Networks (CNNs) have played an essential role in feature extraction throughout the previous decade [9]. Local features in CNN are basically the feature maps derived from the network intermediate convolutional layers, whereas global features are those created by the whole CNN architecture. Therefore, fully-connected layers are often supplied with global features as input. However, while the CNN performs admirably in many applications, it has several surprising limits. The CNN limitations include the fact that it can only provide image predictions if and only if the compared images are almost perfectly aligned. Also, it lacks encoding for image posture and orientation.

Furthermore, a CNN does not examine the relative positions of the traits in relation to one another. It also transports high-dimensional data from lower to higher levels. Indeed, rather than propagating images (i.e., information) via all neurons, it is preferable to direct them to specific neurons that can cope with certain properties. This process resembles the human brain actions, in which distinct parts decode different types of information. Doing the same in a Neural Network (NN) improves predictions [10], [11], [12].

Nonetheless, applications that rely on visual localization analysis do not reach the same degree of posture precision. In addition, existing techniques do not consistently beat a hand-made image retrieval baseline [13]. Therefore, it should be highlighted that much effort is being made to rectify CNN shortcomings. However, the baseline for the whole point will be determined by thoroughly understanding the identified feature uniqueness and discrimination ability. This will aid in the decision-making process for the optimum feature extraction algorithms.

One of the challenges for the CBIR system to work over large-scale datasets is labeling of training data. The generalizability of the learned deep representations of new classes is thus constrained by the requirement of supervised training for all target images [24]. Insights are therefore directed to either unsupervised or semi-supervised learning approaches. Unsupervised learning is the general problem of extracting useful information from enormous volumes of unlabeled data. Thus, convolutional autoencoders, Information-Maximizing Generative Adversarial Networks (InfoGAN), and Vision Transformer (ViT) are utilized as unsupervised learning methods for CBIR.

Most CBIR systems use reduction or selection for derived feature descriptors to have low-dimensional vectors in order to speed up similarity matching and retrieval. Convolutional autoencoders will be used to retrieve images after compression, while reducing the number of parameters required. Autoencoders are used to extract features, since it is desirable to represent images with low-dimensional features [14].

In this case, the created latent vector serves as a feature descriptor, representing the content of the image in the feature space. It is worth noting that in terms of dimensionality reduction, the autoencoder is comparable to Principal Component Analysis (PCA), but it is more powerful and intelligent.

In addition, Generative Adversarial Networks (GANs) are computational structures that generate new, synthetic examples of real data. They are commonly utilized in image, video, speech, and sketch retrieval. Information Maximizing Generative Adversarial Networks (InfoGANs) are other generative adversarial networks that maximize the mutual information between latent variables and the observation. They are also used in extracting features for content-based retrieval through the knowledge of the InfoGAN discriminator model [15]. The transformer is a brand-new type of neural network that extracts intrinsic features through the self-attention process, and it has much potential for Artificial Intelligence (AI) applications [16], [17]. Transformers are employed, because they relate various positions in the same sequence, using the self-attention approach to give a representation of the features within images.

B. MOTIVATIONS

Image retrieval system is a feature-based system that is the key engine of various disciplines, such as medical applications, web browsing, satellite imaging, and others. Moreover, the need for such a system has become inevitable with the proliferation of touchscreen devices and the expansion of large-scale web browsing. The inability to retrieve images depending on the user's intent is a roadblock to meeting his needs. This depends upon the similarity matching procedure, which examines the spatial distance between generated features from images to delete irrelevant images. It establishes the optimal correspondence of image distinctive features to those of other images by assessing the similarity of feature details between a query image and a set of other images [1], [2].

C. LIST OF CONTRIBUTIONS

This research is focused on evaluating the effectiveness of image retrieval using various deep feature extraction algorithms. As a result, the following items are the paper primary contributions:

1) Facial sketched-to-real image retrieval is difficult to address since faces are made up of different sections, and images from two different domains are compared. Therefore, in this paper, a novel dataset entitled ESRIR is created by a number of augmentation techniques for the CUFS dataset in order to address the need for large-scale data for face retrieval.
2) For retrieving ESRIR queries, a novel architecture is proposed adopting three different image retrieval systems based on convolutional autoencoder, InfoGAN, and ViT, under changing viewpoints of visual scenes.
3) The three suggested system designs are trained on two additional facial datasets of different image forms due to the prevalence of multiple query forms, particularly with the rise of touchscreen devices.
4) The proposed architectures illustrate that the network power can be used to boost an existing system performance for image retrieval. As a result, the retrieval performance of such systems is also evaluated and practically compared to those of cutting-edge methods.
5) The suggested systems are tested on a challenging dataset of doodle-scratched human artworks.
6) A multi-category dataset is also examined to cover almost all potential image types and circumstances.
7) Finally, systems that use logo content pictures for brand logo search, another widely used image search type, are investigated. Therefore, using visual features extracted from each of the three suggested retrieval systems, we compare the performance of logo image retrieval.
8) The paper also presents cloud-based energy and computational complexity saving approaches for large-scale datasets.

D. PAPER STRUCTURE

The following sections make up the structure of the paper. Section II is focused on the related work with a discussion of the current feature extraction and retrieval methodologies. The definition of the involved problem is shown in Section III. The suggested image retrieval system designs are shown in Section IV. Section V gives a discussion of how image retrieval performance can be tested. The utilized datasets are explored in Section VI. A summary of the involved test scenarios is provided in Section VII. The performance comparison of all trained models is introduced in Section VIII. Section IX describes the experimental settings as well as the results obtained. The analysis of outcomes is given in Section X. Finally, section XI gives the conclusion.

II. RELATED WORK

One of the most challenging problems is the sketch-to-real matching used in CBIR applications. Sketched-real retrieval is often conducted using object edge features and other detected features. It is rare to use learned features in real images, when compared to sketched images. Either handcrafted features or learning algorithms are used. Features detected from the query sketch are compared to those collected from real images. Several studies have been presented for image retrieval inside sketched-drawn representations and their relations to real scenes. Handcrafted features are either global or local. Global features are used to reflect the full image contour. On the other hand, a local feature representation is a pattern or unique structure detected in an image, such as a point, an edge, or a tiny picture patch. The search methods depict the contents by concentrating on a few key areas within the image that vary from region to region and are unaffected by variations in perspective or illumination.

The tensor-based image descriptor, which extracts global features for the edges and outperforms the edge histogram descriptor, was presented in [18]. However, the global visual features may not work effectively, when the target images have several background clutters. As a result, they might be used as a supplement to improve image retrieval accuracy [3]. In addition, to solve the shape-to-image matching problem, the Angular Radial Partitioning (ARP) technique was introduced [19]. A mixture of angular and radial partitioning is used to improve angular partitioning. In the ARP, the image is segmented into $M \times N$ sectors when the edges are detected, where M represents the number of partitioning angles and N is the number of radial partitioning angles. However, the ARP is vulnerable to affine transformations, which impede image matching and lower the retrieval performance.

Furthermore, data retrieval speed and system performance are degraded as computational complexity rises. Therefore, the Angular Radial Orientation Partitioning (AROP) technique, which employs global and local features in the matching process, was proposed [20]. It uses two types of image contour maps: global contour maps and salient contour maps. The Berkeley detector is used to get the contour maps, and Regional Contrast (RC) extracts the salient image regions from dataset images. The AROP features are then specified using the retrieved candidate contour maps. Indeed, the newly disclosed AROP feature methodology is an enhancement of the ARP method by using orientation partitioning. As a result, the AROP features have total dimensions of $M \times N \times O$. Thus, each sector is represented by the number of pixels under different orientation maps by the AROP feature. Although the AROP technique is orientation-invariant, it still has scaling and translation dependencies. Furthermore, its computational cost is relatively high, which slows down the matching process. In addition, the authors of [21] described the Edgel (edge pixel) index approach for pixel-to-pixel matching, where the shape-to-image matching challenge was solved using the local features technique. A mind finder is a real-time image retrieval system that matches pixels at the pixel level. Its purpose was to deal with the problem of Sketched-Based Image Retrieval (SBIR) on a large scale. For contour comparison, distance maps are constructed using Oriented Chamfer Matching (OCM) as a similarity measure. The hit maps, which are binary similarity maps used to construct the Edgel Index Structure (EIS), are then transformed from these maps. However, due to the high computational cost, the Edgel approach is unreliable, when dealing with local affine fluctuations.

The bag-of-features technique [22] was proposed to address the shape-to-image matching challenge with local features. These local features are extracted using the Canny edge detector and the Scale-Invariant Feature Transform (SIFT) descriptor. The bag-of-features technique has outperformed standard global descriptors, but with a high computational cost. SIFT enforcement is not the ideal solution owing to the sparse spatial distribution of its detected points and the large dimensions of its calculated descriptors [23]. In addition, it is inadequate for large-scale datasets.

In [24], the authors proposed TOP-SIFTs, a descriptor selection approach based on dictionary learning, to eliminate duplicate features. Dictionary learning, which works with sparse data, is reserved for a few excellent geographic distribution features. There are two practical shortages that result from this approach. The first is with SIFT, which is technically and computationally challenging. The second point of concern is the selection method, because the whole descriptor computation is required to be completed first, followed by the selectivity procedure. Hence, additional computation enumeration is introduced. As a result, similarity matching in any image-retrieval system will be hindered. Besides, several problems emerge in matching on large-scale datasets. Consequently, the authors of [23] proposed matching based on approximate shapes, with the object represented as a collection of recognized approximate forms termed as primitives. Each primitive has various descriptive parameters as well as information about its kind. This method of rapid access exists, but it has only been tested on small datasets. This may not be possible to maintain with the rapid growth of web images.

A paradigm for creating visual representations that capture the scene essential components and semantic notions was proposed in [25]. To create features with semantic associations, first, connections between picture areas are put up, and then the reasoning is accomplished using Graph Convolutional Networks (GCNs). The gate and memory methods are then used to execute global semantic reasoning based on these relationship-enhanced attributes. As a result, discriminative information is selected, and a representation of the entire image is progressively generated.

Since the human face is composed of many different parts, facial retrieval is considered challenging. It is more complicated to distinguish facial sketches and retrieve their relatives from real ones. Therefore, in [26], InfoGAN has been used to extract a human facial image from the real alternative images based on the discriminator learning of image features. However, the comparison is made across comparable feature domains, which could not provide a thorough evaluation of the InfoGAN architecture [26]. The recommended InfoGAN architecture, which is further described in this paper, is an InfoGAN with a deeper learning discriminator model. Its practical applicability is examined, aiming for generality across a range of image types, image contents, and feature domains.

The ''Transformer'' is a neural network that extracts intrinsic features through self-attention. In [27], DeepViT has been introduced. Unlike CNNs, it stacks more convolutional layers and saturates quickly, when grown to be deeper. Furthermore, it has been found that the self-attention mechanism fails to acquire useful ideas for representation learning in higher layers of ViTs, preventing the model from achieving the required performance boost. Hence, re-attention is

given as a simple yet efficient way for re-generating attention maps to improve their variety at different levels with little computing and memory cost. However, the Principal Vision Transformer (PViT) is data-hungry, because the self-attention layer of ViT lacks localized inductive bias. Image pixels are thought to be locally related with feature translation-invariant correlation maps. So, in [28], Shifted Patch Tokenization (SPT) and Locality Self-Attention (LSA) have been established as broad and effective add-on modules that may be used for a range of ViTs with ease. Hence, in this paper, a proposed architecture is described in Section (IV). It depends on this concept for building a complete image retrieval system. It is trained, examined, and assessed over different types of images and feature domains. Besides, a comparison is established with other state-of-the-art approaches.

Certainly, several applications benefit from CNNs. The authors of [29] thoroughly surveyed the relationship between ConvNet and various pre-trained learning models and their optimization algorithms. Besides, they performed experiments on the face and skin detection datasets to strengthen their survey, highlighting the fitness of pre-trained learning for optimized ConvNet. Convolution-Sparse Filter Learning (CSFL), a novel approach for unsupervised learning, has been presented by the authors in [30] as a way to extract detailed and distinct features from an image. In this approach, the first CNN layer is initialized using the features retrieved by the CSFL method, and the CNN then uses these features in a feed-forward fashion to learn high-level features for classification. The output layer of the CNN is the linear regression classifier (softmax classifier), which provides the likelihood of an image class. The authors of [31] used the Modified Resilient back-Propagation (MRPROP) method to increase convergence and effectiveness of CNN training. The global best notion for weight update criteria is combined with a tolerance band, which prevents network overtraining, to enable the CNN training algorithm to optimize weights more quickly and precisely. In [32], the authors discussed the impact of CNN as one of the most prevalent deep learning models for extracting features for image classification. However, the work of [29], [30], [31], and [32] mainly depends on CNN, which has certain drawbacks, since it can only provide image predictions if the compared images are precisely aligned. It lacks encoding for image posture and orientation. Transformers outperform CNN designs, because they entirely avoid recursion by processing sentences and learning word associations using multi-head attention mechanisms and positional embeddings. Thus, this paper introduces an image-retrieval system based on ViT and highlights its achievements compared to CNN models. With the accomplishment of a fast retrieval procedure based on image features with high retrieval accuracy, this paper is a step forward in dealing with various sorts of images over different domains.

In [33], the authors highlighted the importance of CBIR systems. They introduced a CBIR system based on the Deep Search and Rescue (DNN-SAR) algorithm for retrieving relevant images. Their proposed system includes various steps such as pre-processing, multiple feature extraction, feature fusion, clustering, and classification. They used multiple-feature extraction for the fusion of different features into a single vector. Then, clustering is performed using the adaptive Sunflower Optimization (SFO) algorithm. However, this worthwhile effort includes a number of processing steps, such as classification before retrieval, which will affect the pace of retrieval, especially with the explosive growth of online multimedia content.

In [34], the authors introduced a CBIR system using a dense angle descriptor and Dictionary Learning (DL). They proposed a dense angle-based Histogram of Gradients (HoG) descriptor to address image rotation. Their proposed methodology depends on estimating angles across multiple scales and a bag of visual features at different scales. Such algorithm aims to find a frame, where training data allows a sparse representation. Actually, such algorithm could be inadequate for large-scale datasets, as processing enumeration will be introduced. As in the bag of visual features, images are treated as words, and then similarity matching is then applied. As a result, the retrieval process will take longer time.

Several deep-learning-driven algorithms for identifying and creating sketches have been investigated in [35]. Swire, a method for querying massive libraries of design examples with sketches, is the first of two unique technologies introduced. The second is Sketchforme, a system that creates sketched scenes from natural language descriptions provided by the user. In [36], the CUFS dataset was introduced. It has 606 faces commonly used to recognize and synthesize face renderings due to the rarity of large-scale sketched-real datasets and face image datasets. This necessitates the creation of a large-scale dataset of sketched-real facial images. As a result, a new dataset is developed in this study to expand the original CUFS. The use of several augmentation techniques creates a novel one. The performance of the introduced retrieval systems is also evaluated in this research through training using the supplied dataset and user testing of interactive use cases. Our evaluation reveals that these technologies might successfully support interactive applications and provide new avenues for human-computer interaction in the fields of art, education, design, and other areas.

It is worth noting that there are several strategies for generating image sets, as seen in [37], where the authors constructed the FB5K image dataset using Facebook crawling. They introduced a cross-media retrieval system based on Optical Character Recognition (OCR) with the incorporation of high-level semantic information. However, this valuable work of web crawling will be inconvenient for the proposed facial image retrieval in this research. This is due to several

issues. First, the main goal is to create a large-scale image dataset with various groupings of images for each individual. Second, it is hard to find groups of images for the same person of these sorts (i.e., in real and sketched forms) of images. Third, developing and examining retrieval systems is required to learn and return similar images at different positions and scenes. Thus, augmentation is utilized to generate the proposed ESRIR dataset.

III. PROBLEM DEFINITION

Generally, a new sort of search, image search, has been developed recently to meet users' goals and it might be used instead of or in addition to the more usual language-based image retrieval (e.g., Google image search). In addition, several query types are accessible, including query by example image, query by sketch map, query by color map, query by context map, etc. In the computer graphics industry, recent applications such as CBIR and SBIR that are feature-based are used for these types of search. For real-to-real image matching, various challenges from quantum standpoint, quality, and type of images perspectives confront any image search engine or retrieval approach. In addition, the quality of the query image heavily influences the visual cues that differentiate objects and shapes inside images. To satisfy the user's subjective requirements, a compromise is made between the demands of the system and the query quality employed.

Sketches are free-hand drawings without color or substance. Sketched-to-real matching, or Sketch2Photo, is widely used in various applications as an image synthesis approach. It is also good in sketch-based image retrieval. Sketched-real retrieval is one of the most difficult issues to tackle. The issue is that images from two different domains must be compared, and it is difficult to distinguish them from other types of imagery. This is attributed to the fact that differentiating objects, features, and humans in sketched representations and recognizing their counterparts in real images is difficult due to changes in information type and size within images. Retrieving human faces increases the distress of sketched-real image similarity matching challenges as faces have different parts to be recognized. Hence, objects or humans' peculiar feature separation from all derived image features is a noteworthy difficulty with these sorts of images.

Any retrieval system must address three aspects: average retrieval accuracy, retrieval speed, and memory cost. The content of images increases the complexity of these systems. Due to the large number of image features, the retrieval system capacity to match images quickly will be affected. This draws attention to the necessity for efficient and appropriate representation of image content. In addition, as the volume of multimedia data grows, feature extraction algorithms face a problem in meeting the demands of these systems. Thus, object features must be bundled into convenient dimension descriptors to speed up image matching. This isolates the most object discriminative features, independent of image redundancy, to distinguish humans or objects within images.

As a result, the ability and robustness of extraction methods to properly describe picture contents with adequate descriptors to boost performance accuracy is the main engine of the entire process. Deep Learning (DL) networks are frequently used to extract features and reduce the data dimensionality. CNN is the most widely-used DL algorithm. It uses convolutional layers and pooling layers for processing of shift-invariant inputs such as images. Hence, it is a type of deep neural network that is frequently used to evaluate visual images. Instead of manually implementing the feature extraction, CNN performs it in the training phase. The CNN feature extractor is made up of many types of neural networks that determine the weights throughout the training phase.

The CBIR may be thought of as an unsupervised learning method based on DL. Autoencoders are unsupervised neural network models that learn how to recover data after compression, while summarizing its general features in fewer parameters. An autoencoder is similar to PCA in terms of dimensionality reduction, but it is more powerful and smarter. As shorter descriptors are required in any retrieval system, autoencoders could be applied in such systems where no class labels are used in training. For such systems, the autoencoder is used to construct the feature vector for each image in the dataset. The feature vector will be the latent-space representation derived by the autoencoder. Then, at the time of search, the distance between the latent-space vectors is calculated. The smaller the distance, the more relevant or visually comparable the two images are.

Because the important downstream tasks are unknown at training time, unsupervised learning is inappropriate. However, a disentangled representation that explicitly expresses the salient features of a data instance should be useful for the relevant but unknown tasks. For example, each of the following attributes may be assigned a separate set of dimensions in a useable disentangled representation for a dataset of faces: facial expression, eye color, hairstyle, presence or absence of glasses, and connected person's identity. Natural tasks that require knowledge of the data prominent qualities, such as face identification and object recognition, can benefit from a disentangled representation. Generative modeling drives a substantial portion of the unsupervised learning research. It is motivated by the belief that synthesizing or creating observed data entails some level of understanding. It is hoped that a good generative model will learn a disentangled representation automatically, even though perfect generative models with arbitrarily-bad representations are easy to construct. InfoGAN could be deployed for facial image retrieval. It is an adversarial generative network that maximizes the mutual information between observation and selection of latent variables.

Recurrent Neural Networks (RNNs) handle sequential or time series data using recurrent cells. They are sequential in processing. Therefore, the previously calculated hidden states of the first pixel are required to encode the second pixel in the image patch. The transformer is a brand-new type of neural network with a lot of potential for AI applications.

Thus, it is used for image-retrieval systems, primarily using the self-attention process to extract intrinsic features. In the domain of machine translation, transformers were invented to avoid recursion and allow parallel computation to decrease the training time and performance losses due to lengthy dependencies.

To begin with, a transformer is non-sequential, meaning that it processes data rather than pixels one by one, avoiding long-dependence concerns. Transformers compute similarity scores between pixels in an image patch using the newly-introduced unit ''Self Attention''. Furthermore, to prevent a recurrence, positional embeddings are added. The idea is to encode information about a token location in an image patch using fixed or learned weights.

This paper introduces image retrieval systems based on the computed descriptors from three different learning feature extraction algorithms. It is required to differentiate between them and easily select the convenient one to reduce the gap between the performance of these systems and the requirements of retrieval applications. These introduced systems depend on convolutional autoencoders, InfoGANs, and ViT, respectively. Efficiency has been proven through training and detailed assessment of the systems on various datasets and from various perspectives. This is conducted through the measurement of similarity matching and retrieval between calculated descriptors by each system. Several experimental cases are included in this paper, each of which works on a distinct dataset and an aspect of performance evaluation. Image retrieval performance is compared to those of CNN models as the benchmarks for feature extraction algorithms.

IV. PROPOSED MODELS

This paper presents three architectures to retrieve similar images from a group of images based on the spatial distance between their derived features. The following is a description of the three architectures.

A. 1st MODEL: IMAGE RETRIEVAL BASED ON CONVOLUTIONAL AUTOENCODER

In machine learning, an autoencoder is an unsupervised learning tool for which the input and output values are the same, intending to transform the input into output. It is used to compress data to save storage and reduce processing time by eliminating unnecessary variables, displaying high-dimensional data, and removing noise from the original data [38], [39], [40]. It is based primarily on using neural networks to implement compression and decompression functions. Instead of being carried out by humans, these functions are data-specific, lossy, and automatically learned from examples.

The proposed architecture, shown in Figure 1, reveals that the CNN power can be used to boost the performance of a simple autoencoder for image retrieval. As the figure shows, the encoder and decoder components of the convolutional autoencoder are essential. The encoder is made up of several diminishingly concealed layers. The latent vector, which is used as a feature descriptor for the image, is reached through these layers. As a result, the encoder aspires to learn an image hidden and condensed depiction as an image feature representation. The decoder reconstructs the images using the feature descriptors provided by the encoder. It works in the opposite direction of the encoder. These attributes are then combined with the nearest-neighbor algorithm to determine comparable, and similar images.

The advantage of this process is that image features can be efficiently represented in low-dimensional vectors. Furthermore, comparing image features is preferred to comparing images in their raw form. The convolutional encoder must emphasize certain constraints, and the decoder must produce the input image shape. Besides, it must be trained across all image categories for high retrieval accuracy.

B. 2nd MODEL: IMAGE RETRIEVAL BASED ON INFORMATION MAXIMIZING GANs (InfoGANs)

The GAN is a training architecture for deep convolutional models that generate synthetic images. InfoGAN is a GAN extension that introduces control variables, such as style, thickness, and type. It comprises three sub-models, as shown in Figure 2. The generator model oversees generated images during training, and the discriminator model learns to categorize images as real (from the training dataset) or fake. The third model is an auxiliary model, which can forecast continuous control variables using a Gaussian distribution. Thus, the two models (i.e., the generator and discriminator models) compete in a zero-sum game to find a balance between the generator ability to create convincing images and the discriminator ability to recognize them, which is critical for the training process convergence. The third model, the auxiliary model that predicts the control variables, is created by including control variables. This model is then trained using a mutual information loss function to control which image the model generates.

The generator model takes random points as input from the latent space, gives those points a unique meaning through training, and then maps the points to different output synthetic pictures. While the generator model structures the latent space, there is no control over the resulting image. The latent space may be examined, and the generated images may be compared to comprehend the mapping that the generator model has learned. Alternatively, the generation process can be conditioned, for example, by using a class label, so that images of a given kind can be generated on demand. This refers to the Conditional Generative Adversarial Network (CGAN). Another option is to give the generator control variables as well as the latent space points as input (noise). Control variables can be used to teach the generator to impact certain aspects of the generated images. The InfoGAN has been introduced in [26]. The generator receives control variables as well as noise as input, and the model is trained using a mutual information loss function. The use of a new

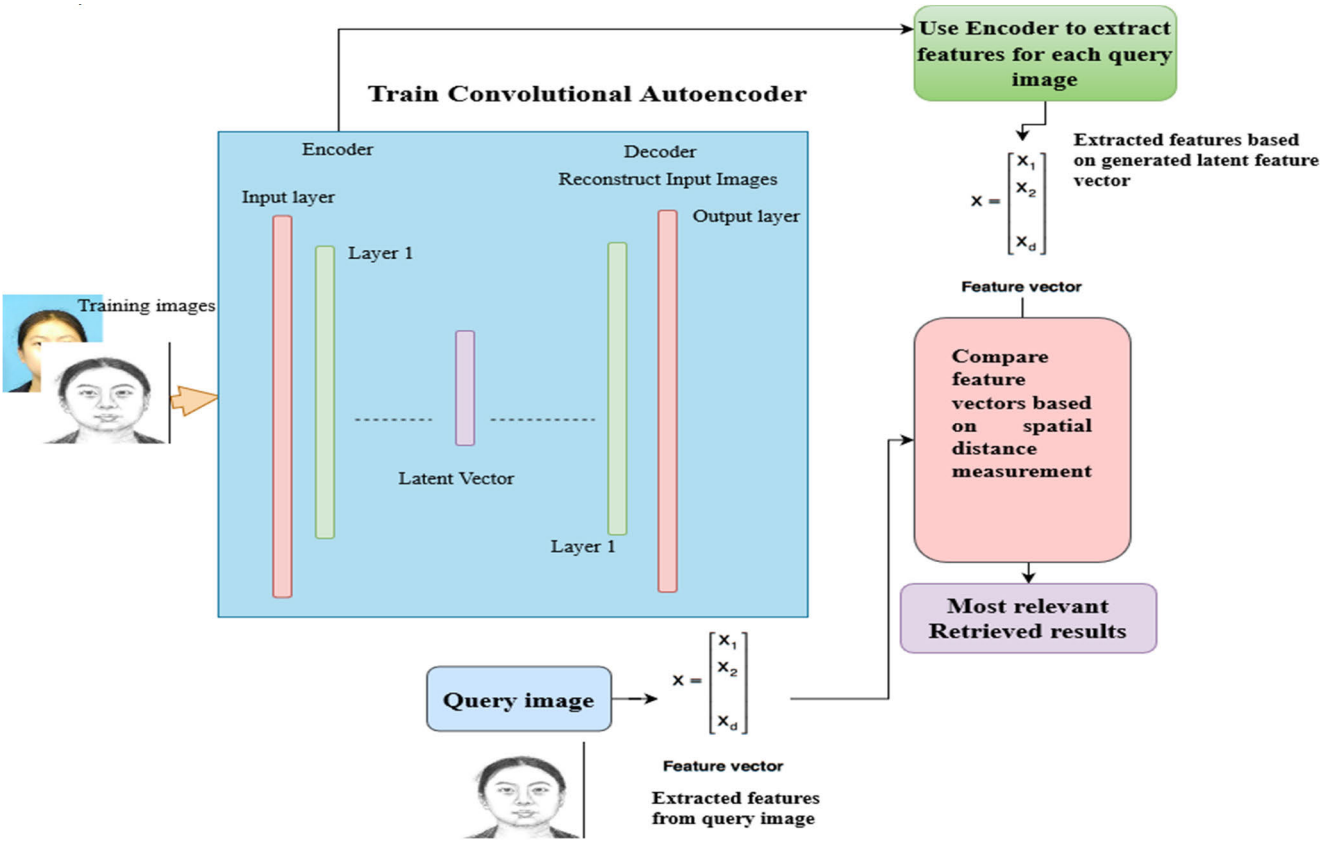


FIGURE 1. First model: Image retrieval using convolutional autoencoder.

model referred to as the ''auxiliary model'' is used to train the generator via mutual information. For interpreting an input image, the new model uses the same weights as those of the discriminator model, but instead of predicting whether the image is real or false, the auxiliary model predicts the control codes used to create the image. Thus, both models are used to update the generator model. The first enhances the likelihood of generating images that deceive the discriminator model, and the second improves the mutual information between the control codes used to produce an image and the control codes predicted by the auxiliary model. Due to the mutual information loss, the generator model is regularized, so that the control codes may effectively govern the image generation process and capture the main aspects of the created images.

Once the InfoGAN has been trained, the discriminator may be used to search for images that are similar to one another. The notion is that the network learns meaningful features from images based on mutual knowledge, such as the Physiognomy of individuals in a picture, as Figure 2 shows. Thus, the network gives a sufficient and good feature representation for images. Moreover, indexing is carried out to investigate the output features. Finally, image similarity is estimated using a spatial distance measure between derived features.

C. 3rd MODEL: IMAGE RETRIEVAL BASED ON VISION TRANSFORMER

Attention is a phrase used to describe the relationships between pairs of input tokens. The cost grows exponentially as the quantity of tokens increases. Transformers measure these relationships depending on pixels as the basic unit of image analysis. Although transformer architecture has established the de facto standard for natural language processing, its applications in computer vision are still limited. It considers an input image as a series of patches, like how a Natural Language Processing (NLP) transformer generates a series of word embeddings. The transformer takes a list of words as input and uses them for categorization, translation, and other NLP tasks [35]. On the other hand, in vision, attention is used either in combination with convolutional networks or to replace components of convolutional networks, while maintaining their general structure. The suggested design guarantees, however, that the use of CNNs is not required, and that a pure transformer applied directly to a sequence of image patches may get outstanding results on image retrieval tasks.

Because the self-attention layer of the ViT lacks localized inductive bias, the major ViT difficulty is data hunger. Those image pixels are locally connected and have translation-invariant correlation maps. Therefore, ViTs require addi-

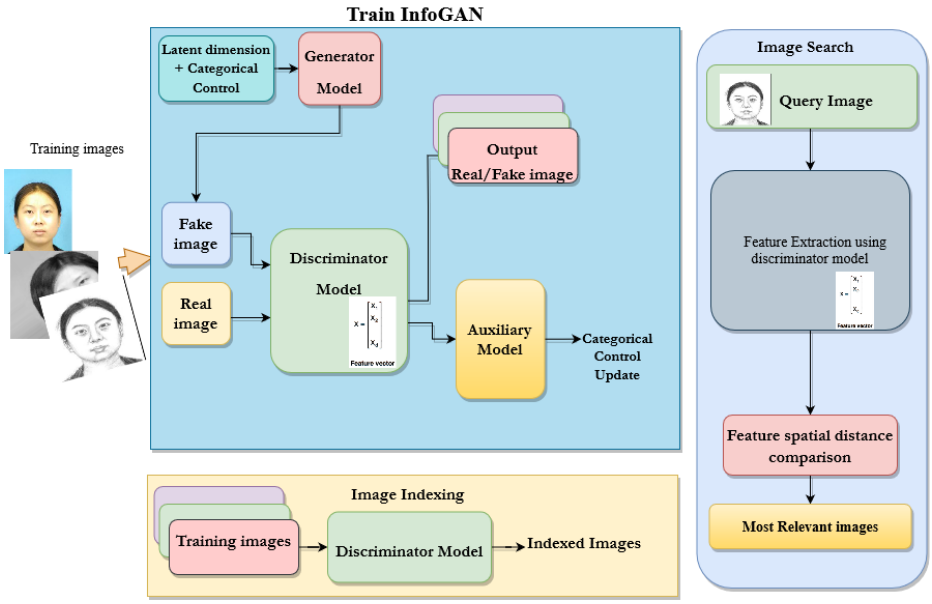


FIGURE 2. Second model: Image retrieval based on InfoGANs.

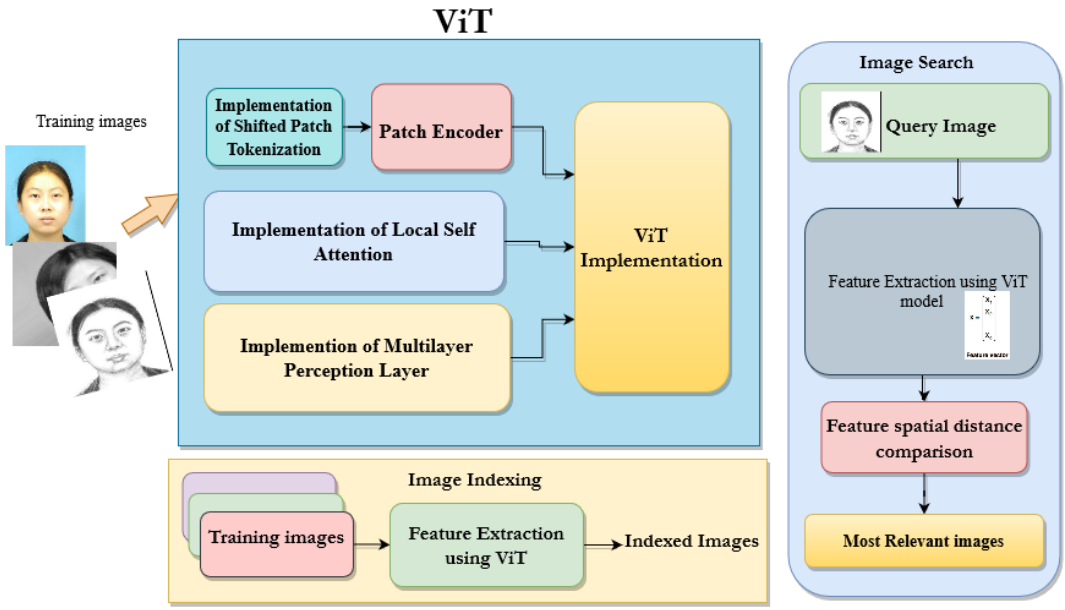


FIGURE 3. Third model: Image retrieval using ViT.

tional information. CNNs, on the other hand, examine images through spatial sliding windows, allowing them to achieve greater outcomes with fewer datasets. As a result, the only way to outperform state-of-the-art CNN models is to pretrain a ViT on a large dataset and then fine-tune it on medium-sized datasets. The suggested architecture depends on ViT, which was introduced for small datasets. The engaged ViTs overcame the absence of localized inductive bias and are trained from the beginning, even on small datasets. Shifted Patch Tokenization (SPT) and Locality Self-Attention (LSA) have been introduced as general and effective add-on modules that may be simply applied on a variety of ViTs [28]. For retrieval tasks based on ViT, the proposed architecture depends on ViT descriptor for image representation through the features learned by ViT itself, unlike the work introduced in [41]. Thus, image representation is implemented with ViT through SPT implementation and then encoding of patches. After that, the most important step is implementing LSA. Then, images are indexed according to the feature descriptions, and a search is performed according to the spatial distance between such features, as Figure 3 shows.

V. IMAGE RETRIEVAL PERFORMANCE CRITERIA

Several ways of assessing the performance of a system have been developed in the literature. Because Information Retrieval (IR) has been used to solve many problems, the relationship between CBIR and IR is evident. Despite the differences across the fields, many of IR solutions may be applied to CBIR. This section gives a discussion of the performance indicators and methods for building a standard test suite for CBIR. One of the most critical and time-consuming duties is determining which images are relevant and which are not for a specific query. This depends mainly on similarity matching as the core engine of the whole procedure.

Finding the best correspondence q points of an image to previously extracted N points of interest of another image [4] is a frequent definition. The spatial distance between descriptors created from images using the deployed feature extraction algorithm is measured. The effectiveness with which a region descriptor represents a scene region is determined by its strength. This is demonstrated by comparing matching scores. The matched regions are the descriptor space closest neighbors. The matching is performed after the descriptor computations by applying the extraction algorithm. Then, spatial distance is measured between the descriptors. Finally, two metrics are computed to determine how well regions are matched: recall and precision. They are two terms that are often used, interchangeably. Precision and recall are numerical values that vary from 0 to 1. The most common evaluation metrics used in image retrieval are precision and recall (see Eqs. (1), (2)), usually presented as a precision versus recall graph (PR graph) [42], [43]. Another indication is the error rate, which is used as a standard metric in object and image recognition (see Eq. (5)).

$$\begin{aligned} &\textit{Recall} \\ &= \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Negative}} \\ &= \frac{\textit{Number of relevant images retrieved}}{\textit{Total number of relevant images in the collection}} \end{aligned} \tag{1}$$

$$\begin{aligned} &\textit{Precision} \\ &= \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Positive}} \end{aligned} \tag{2}$$

$$\begin{aligned} &\textit{Error Rate} \\ &= \frac{\textit{Number of relevant images retrieved}}{\textit{Total number of images retrieved}} \end{aligned} \tag{3}$$

The precision and recall metrics can be merged into a single result known as an F-score, which defines their harmonic mean (see Eq. (4)). The F-score is a single factor of record-matching accuracy that incorporates precision and recall measurements. As a result, it accounts for both false positive and false negative errors.

$$\text{F-score} = 2\frac{\textit{Precision.Recall}}{\textit{Precision} + \textit{Recall}} \tag{4}$$

Another metric is retrieval efficiency, which determines whether the number of images retrieved is less than or equal to the number of relevant images (see Eq. (5)). If the number of images retrieved is less than or equal to the number of relevant images, this value is the precision; otherwise, it is the recall of a query. This definition may be confusing, because it combines two standard measurements [44], [45], [46].

$$\textit{Retrieval Efficiency} = \begin{cases} \dfrac{N_R}{N_T}, & \textit{if } N_T > N_R \\ \dfrac{N_R}{N_{TR}}, & \textit{Otherwise} \end{cases} \tag{5}$$

where N_R is the number of relevant retrieved images, N_T is the total number of retrieved images, and N_{TR} is the total number of relevant images.

Another statistical metric employed in object retrieval measurements is the correct and incorrect detection. The number of correct and incorrect classifications is tracked. These values are like error rate and precision, when divided by the number of recovered images [47]. Thus, for evaluating the model efficiency in extracting features with good spatial and discriminating attributes from images and retrieving them, recall, 1-precision, and F-score are computed in each case. In addition, the computation of other defined metrics leads to the same conclusion. Computational efficiency is also included. It refers to the time cost of visual feature extraction, indexing, and image querying. This, in turn, contributes to the retrieval system efficiency [3].

VI. DATASETS

Six datasets are used for the training of all involved models. First, we present our generated dataset named ESRIR, which depends on the Chinese University of Hong Kong (CUHK) CUFS dataset [36]. The 606 faces in the original CUFS dataset are widely used to recognize and synthesize face drawings. There are 188 students in CUHK. An artist made a sketch for each student based on a photo taken in normal lighting conditions, in a frontal position, and with a neutral expression, as seen in Figure 4. ESRIR is used to bridge the huge domain gap between a facial drawing and a photo. After scraping the original CUFS from the web, ESRIR is created. It is generated through the repetition of images by applying scaling and normalization, only. Besides, augmentation is applied on other images, as rotation is performed on images randomly. In addition, images are randomly translated vertically or horizontally using width and height shifts. In addition, trimming transformations, zooming, and horizontal flipping are used. Finally, the filling produces new pixels that may arise following a rotation or a width/height shifting. As shown in Figure 5, the resulting extended CUFS contains 53,000 sketches and 53,000 images. Very abstract human face sketches, as opposed to the typical semi-photorealistic ones seen in existing datasets, are purposefully sourced to reduce the domain gap.

The second used dataset is the CUHK Face Sketch FERET (CUFSF) dataset, which is also used for face sketch synthesis and face sketch recognition research [48]. It includes 1,194 face images with lighting variations and sketches with shape

FIGURE 4. Original CUFS samples.

FIGURE 5. ESRIR samples.

exaggeration drawn by an artist when viewing the images. It must be noted that the comparison on this dataset is performed between the sketched photos and their cropped ones as sketched-sketched retrieval. The aim of this is to examine models in cropped cases.

The third dataset is Labeled Faces in the Wild home (LFW) [49]. It is a public benchmark for face verification collected from the web, also known as pair matching, consisting of 21,174 facial images. The aim behind the usage of such a dataset is to examine models for real-real retrieval under poor lighting, extreme poses, strong occlusions, and low-resolution circumstances.

The fourth dataset was introduced for Zero-Shot Sketch-Based Image Retrieval (ZS-SBIR). It is entitled QuickDraw-Extended [50]. It consists of 330,000 sketches and 204,000 images spanning 110 categories. This dataset is used to examine models for sketched-real object image retrieval.

The fifth dataset is 256_Object Categories [51], a collection of 256 item categories with 30607 images in total. To achieve this, a set of item categories was chosen. Samples were taken from Google images, and any images that do not fit the category were manually removed.

The last dataset is the Flickr Logos 27 [52], collected from the Flickr group. It comprises around 4,000 classes in total, corresponding to 27 logo classes/brands (30 images for each class). It also includes a distractor set of 4207 logo images/classes, the majority of which feature clean logos. Every image in the distractor set has its own logo class.

VII. EXPERIMENTAL SETUP

This research introduces an assessment of image retrieval performance using features obtained from several models applied to different types of images. The introduced assessment is divided into multiple test cases, each assessing retrieval performance for a distinct model based on a specific image type. As shown in Figure 6, this section gives an explanation of the overall procedural flow followed throughout the simulation tests. The first four predetermined datasets in section (VI), named ESRIR, CUFSF, LFW, and QuickDraw-Extended, are utilized in the experiments offered. The following are the steps of experiments for each dataset:

1. Each dataset is split into training, test, and validation sets.
2. Using the training set from each dataset, the CNN models (Inception, Mobilenet, Resnet, VGG16, VGG19, and Xception) are independently trained.

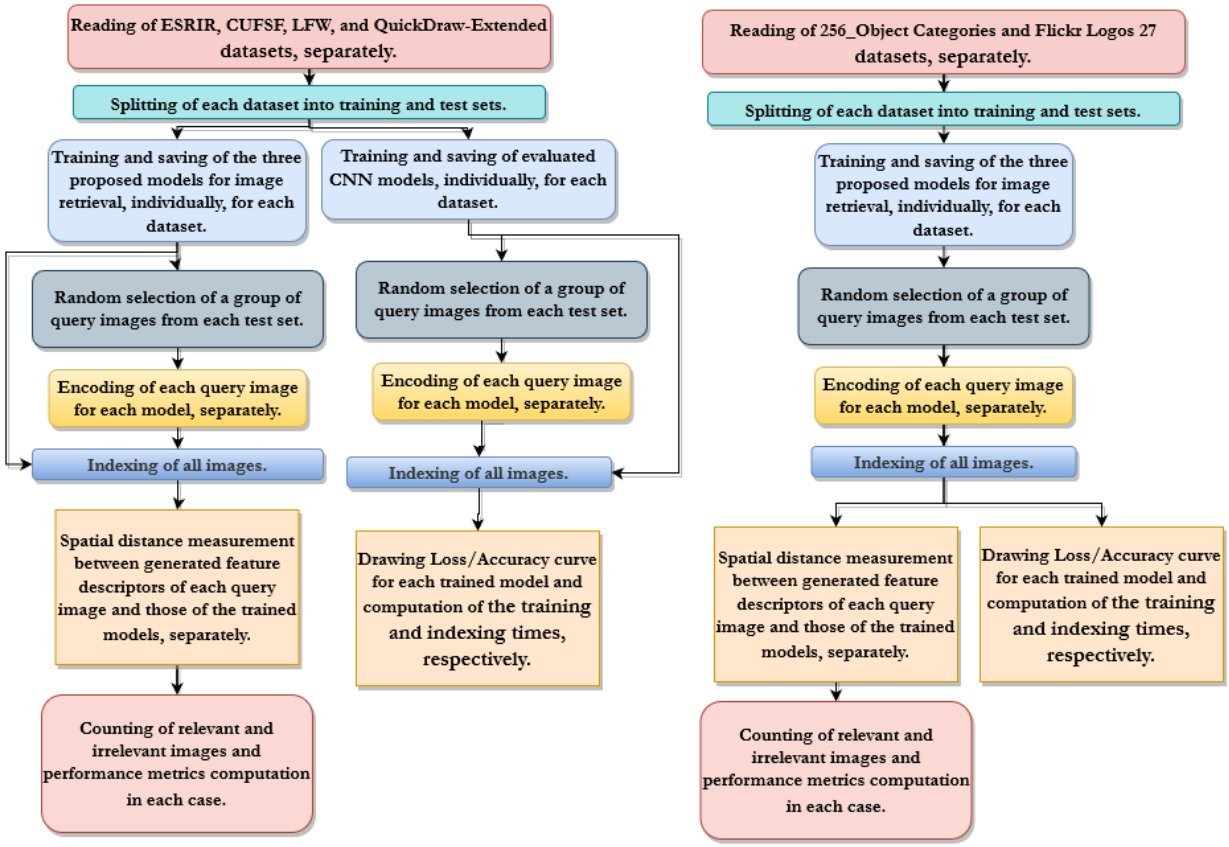


FIGURE 6. General flow diagram for involved test case scenarios.

3. In addition, the three predefined image retrieval models in Section (IV) are applied and trained on each image in the training set of the four datasets. Generally, either CNN models or the three proposed image retrieval systems are trained by the training set of each of the four datasets.
4. After each model has been trained using all four datasets, each trained model is saved.
5. Each model independently encodes a group of randomly-selected query images from the split test set. As a result, each dataset consists of a set of training images used to train the models as well as a set of reference query images encoded by each model.
6. The descriptor of each image is calculated, and it represents the contents of the image as a set of significant features and vectors.
7. The indexing of all images is complete.
8. The spatial distance between the extracted features for each encoded query image and those taught by each trained model is then measured for similarity matching. For each query-train pair, the spatial distance is measured using the distance metric to determine the closest neighbor match in the feature space. Finally, the matched images are retrieved.
9. The correct and false matches are counted at the conclusion.
10. Finally, the matching performance is assessed through the computation of recall, 1-precision, and F-score values using Eqs. 1, 2, and 4. This strategy allows to compare all features (regions) from each training image with those from the query image.

It must be noted that the performance levels of these CNN models are compared to those of the architecture models proposed in Section (IV), as illustrated in Section (VIII).

The last two predefined datasets, named 256_Object Categories and Flickr Logos 27, are used to train the three proposed models in Section (IV), as shown in Figure 6. Following the same procedure, each dataset is split into training, testing, and validation sets. The three models are trained and stored individually by the training set of each dataset. Then, a group of randomly-selected query images from the split test set is encoded separately by each model. Image indexing is performed to retrieve the relevant images. Next, spatial distance is measured for similarity matching between each query image encoded features and the others learned by each model. The whole set of features of each training image are compared to the features of the corresponding query image, in each case.

The calculated query image descriptor is matched with its counterparts (i.e., query-train pair). The distance metric is used to identify the nearest neighbor match in the feature

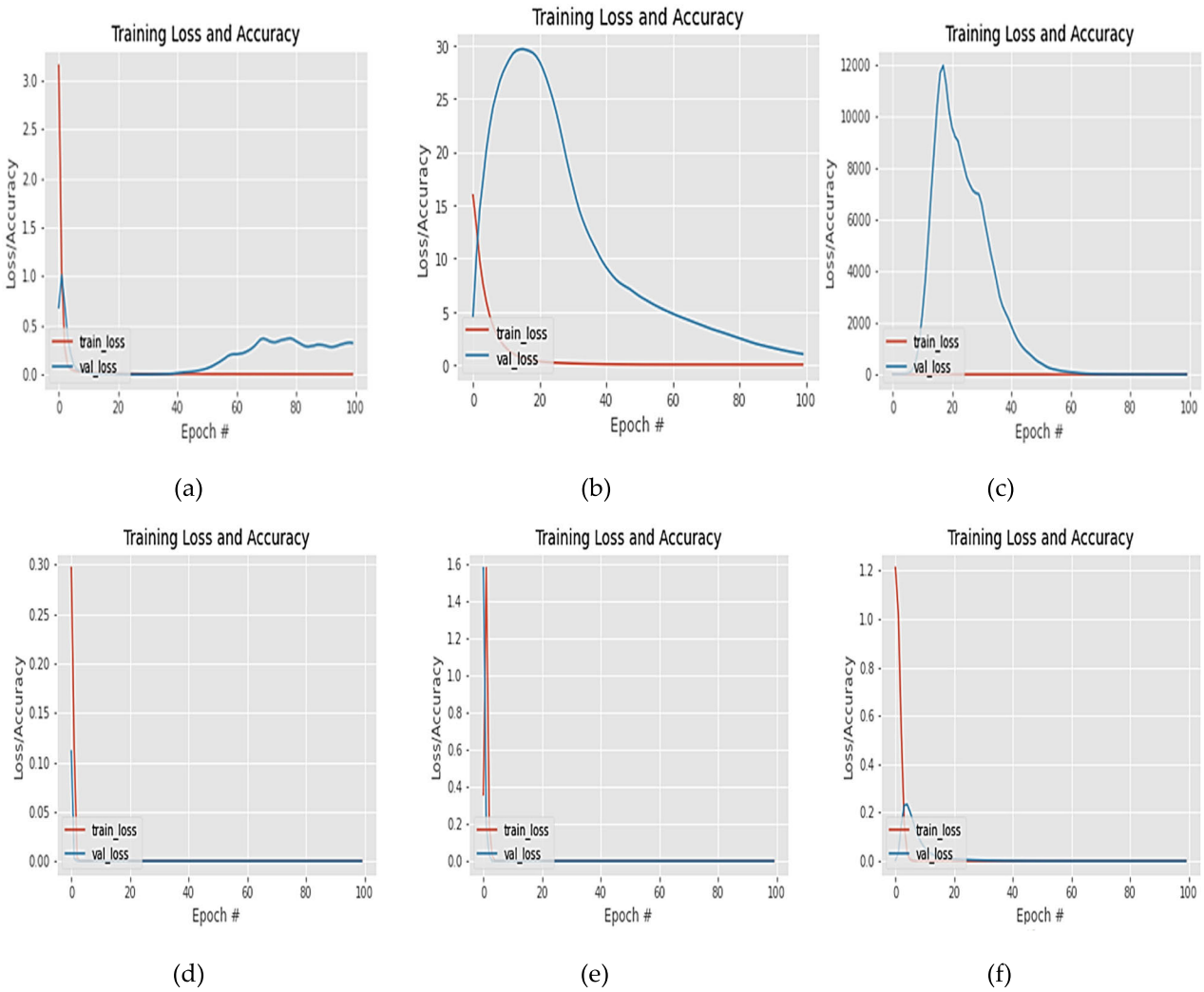


FIGURE 7. Training and validation losses for (a) Inception, (b) Mobilenet, (c) Resnet, (d) VGG16, (e) VGG19, and (f) Xception applied to ESRIR.

space. The closest matched images are returned (the number of correct matches), and the number of falsely-matched images is estimated. Finally, the matching performance is evaluated using Eqs. 1, 2, and 4 to compute recall, 1-precision, and F-score values. It is important to note that the Loss/Accuracy curve is constructed for each model trained on each of the six preset datasets in Section (VI) to provide a comprehensive evaluation of all models involved.

VIII. TRAINING AND EVALUATION MODELS

When creating and configuring DL models, many decisions must be made. Many of these choices may be made by emulating the structure of other networks and using heuristics. In addition, the most effective method is to conduct small tests and scientifically assess possibilities using real data. Thus, this section illustrates performance evaluation comparisons for the three proposed models over all used datasets. Learning curves are a common machine learning diagnostic tool for assessing model learning performance. They may be used to detect learning issues during training, such as underfit or overfit models. The train learning curve is such a curve that is calculated from the training dataset. It indicates how effectively the model is learning, as shown in Figures 7-15. These graphs show the curve, when using the ESRIR, CUFSF, LFW, and QuickDraw-Extended datasets to train models.

- For the ESRIR dataset, Figure 7 shows the loss curve obtained after training of CNN networks (i.e., Inception, Mobilenet, Resnet, VGG16, VGG19, and Xception) using the ESRIR dataset. Figures 11.a, 12.a, and 13.a are the loss curves obtained after training of the three proposed systems based on convolutional autoencoder, InfoGAN, and ViT, respectively.
- For CUFSF dataset, the loss curves produced after training of CNN networks on the CUFSF dataset are shown in Figure 8. These CNN networks include Inception, Mobilenet, Resnet, VGG16, VGG19, and Xception. The loss curves for each of the three suggested systems, which are based on the convolutional autoencoder, InfoGAN, and ViT, are shown in Figures 11.b, 12.b, and 13.b, respectively.

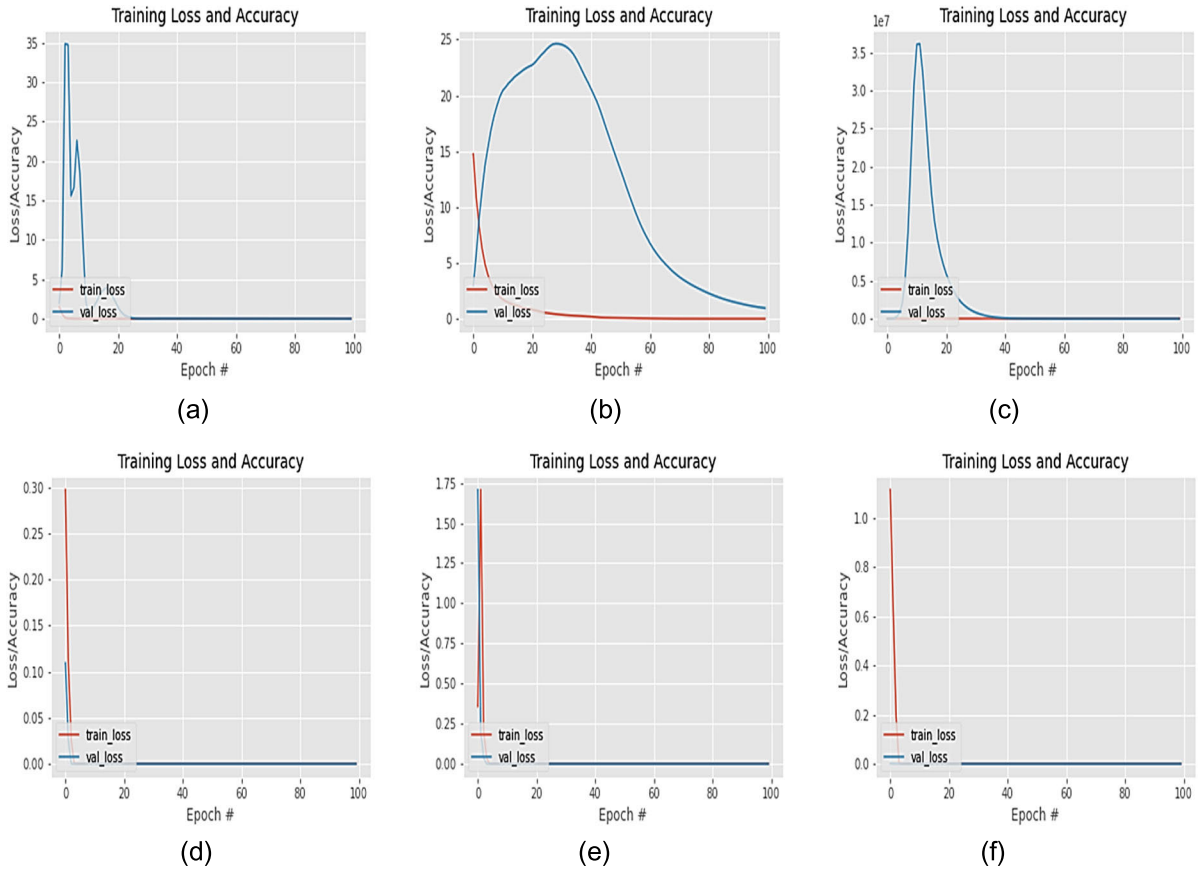


FIGURE 8. Training and validation losses for (a) Inception, (b) Mobilenet, (c) Resnet, (d) VGG16, (e) VGG19, and (f) Xception applied to CUFSF dataset.

- For LFW dataset, Figure 9 displays the loss curves created after CNN networks (i.e., Inception, Mobilenet, Resnet, VGG16, VGG19, and Xception) were trained on the LFW dataset. The loss curves created after training of the three proposed models are displayed in Figures 11.c, 12.c, and 13.c, respectively.
- For QuickDraw-Extended dataset, after CNN networks (such as Inception, Mobilenet, Resnet, VGG16, VGG19, and Xception) were trained on the QuickDraw-Extended dataset, the resulting loss curves are shown in Figure 10. Figures 11.d, 12.d, and 13.d, respectively, show the training loss curves for the three proposed models after training.

Thus, the loss curves shown in Figures 7-10 and those shown in Figures 11-15 are used to assess the performance of the CNN models that were used for comparison with the models suggested in Section (IV).

For efficient performance evaluation of these models, it is worth noting that the learning algorithm aims to get a good fit. That is shown by a training and validation loss that gets down to the point of stability with a small gap between the two final loss values. The model goodness of fit describes how well it matches a collection of data. In most cases, the goodness of fit indicators describe the difference between observed and model-predicted values.

- Figure 7 shows a significant disparity between these losses when training of the CNN models using the ESRIR dataset, particularly for Mobilenet and Resnet models.
- Figures 8, 9, and 10 show how the rest of the four datasets can be used to reach the same result.

Hence, when evaluating trained CNNs, almost all models have an unsatisfactory fit, as shown in the figures. More precisely, CNN models such as Mobilenet and Resnet suffer from underfitting because of a huge difference in training and validation loss curves.

The introduced convolutional autoencoder has a good fit for the ESRIR and CUFSF datasets, as shown in Figures 11.a, and 12.b. For the other two datasets, it has unsatisfactory results to some extent, as shown by the difference between the resulting training and validation loss curves, as Figures 11.c, and 12.a, and 11.d show.

As mentioned, Figures 12 and 13 show the learning curves for the proposed systems using InfoGAN and ViT, respectively. By evaluating their curves over the four used datasets, the proposed models virtually achieve a satisfactory fit, as shown in the figures.

Training loss and validation loss over time are two of the most commonly-utilized measurement combinations. According to the definition of the validation loss, it is the

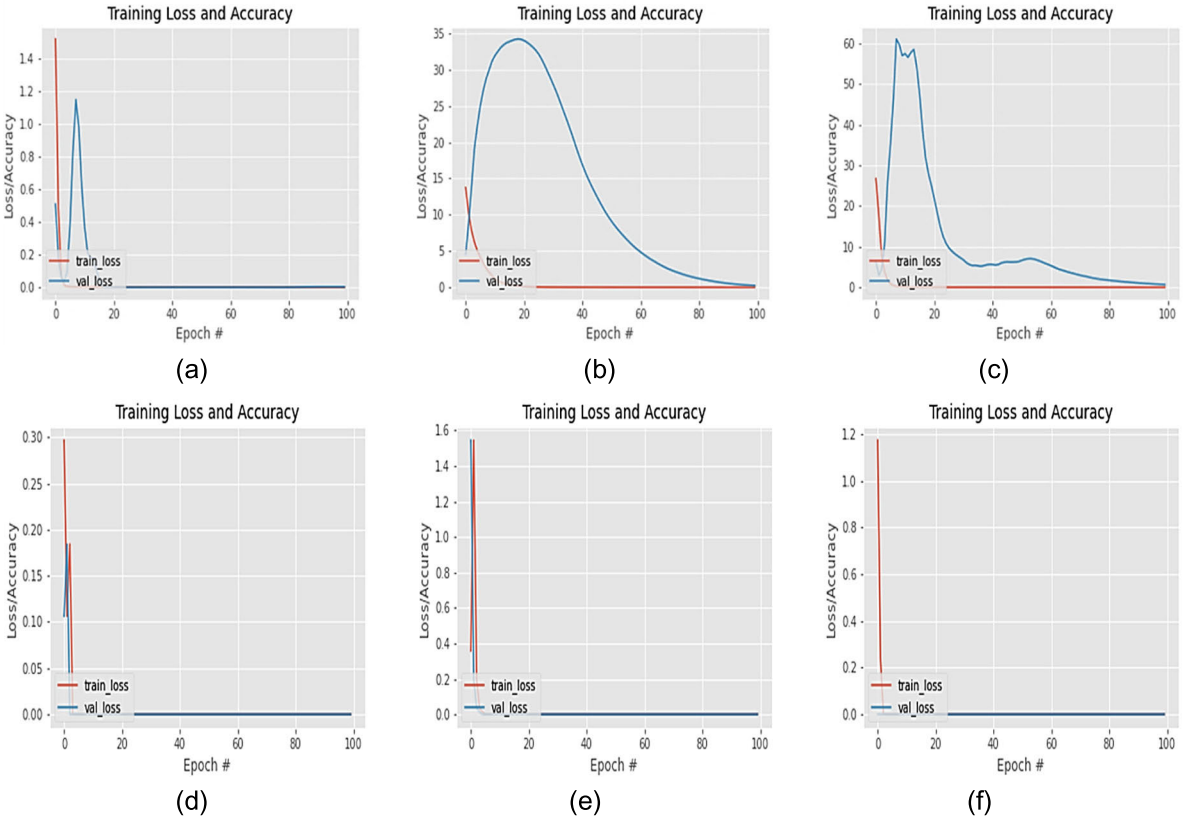


FIGURE 9. Training and validation losses for (a) Inception, (b) Mobilenet, (c) Resnet, (d) VGG16, (e) VGG19, and (f) Xception applied to LFW dataset.

loss estimated on the validation set when the data is separated into train, validation, and test sets using cross-validation. The training loss refers to how well the model fits the training data, whereas the validation loss refers to how well the model fits new data. As a result, the trained model performance could be easily predicted.

- For 256 Object Categories dataset, Figures 14.a, 14.b, and 14.c show the learning curve in the case of training of convolutional autoencoder, InfoGAN, and ViT models, respectively.
- For Flickr Logos 27 dataset, Figure 15.a, 15.b, and 15.c show the learning curves in the case of training of convolutional autoencoder, InfoGAN, and ViT models, respectively.

For the 256 Object Categories dataset, the InfoGAN model is more suitable for the new data than the others, as shown in Figure 14.c. The comparison of both training and validation losses for each model demonstrates this idea. The learning curve for InfoGAN in Figure 14.c is compared to those of its alternatives, the convolutional autoencoder and ViT, shown in Figures 14.a and 14.b. Because the losses do not reach a point of stability, there is a slight gap between the two final loss values.

For the Flickr Logos 27 dataset, it is found that ViT has the best performance among its counterparts, as its loss reaches the stability point faster, as shown in Figure 15.b compared to Figures 15.a and 15.c. It is worth noting that the 256 Object Categories and Flickr Logos 27 datasets were already discussed with CNN. Hence, retraining of them is not necessary.

The training performance of the proposed models will be assessed, as illustrated in the problem definition. It is critical to train and index extracted features to accelerate learning and similarity matching. This verifies the retrieval system scalability across large datasets. As a consequence, Table 1 shows the overall training and indexing times of the models. In he following sections, some ideas based on the findings are introduced.

A. TRAINING COMPLEXITY FOR THE PROPOSED CONVOLUTIONAL AUTOENCODER SYSTEM

It contains 22,030,337 training parameters, as shown in Table 1. The following items could be concluded:

1) TRAINING PARAMETERS

Compared to CNN models, the autoencoder has more training parameters, which are 1.01, 6.82, 1.5, 1.1, and 1.06 times of those of InceptionV3, Mobilenet, VGG16, VGG19, and Xception, respectively. Moreover, those parameters are around 1.071 times fewer than those of Resnet50.

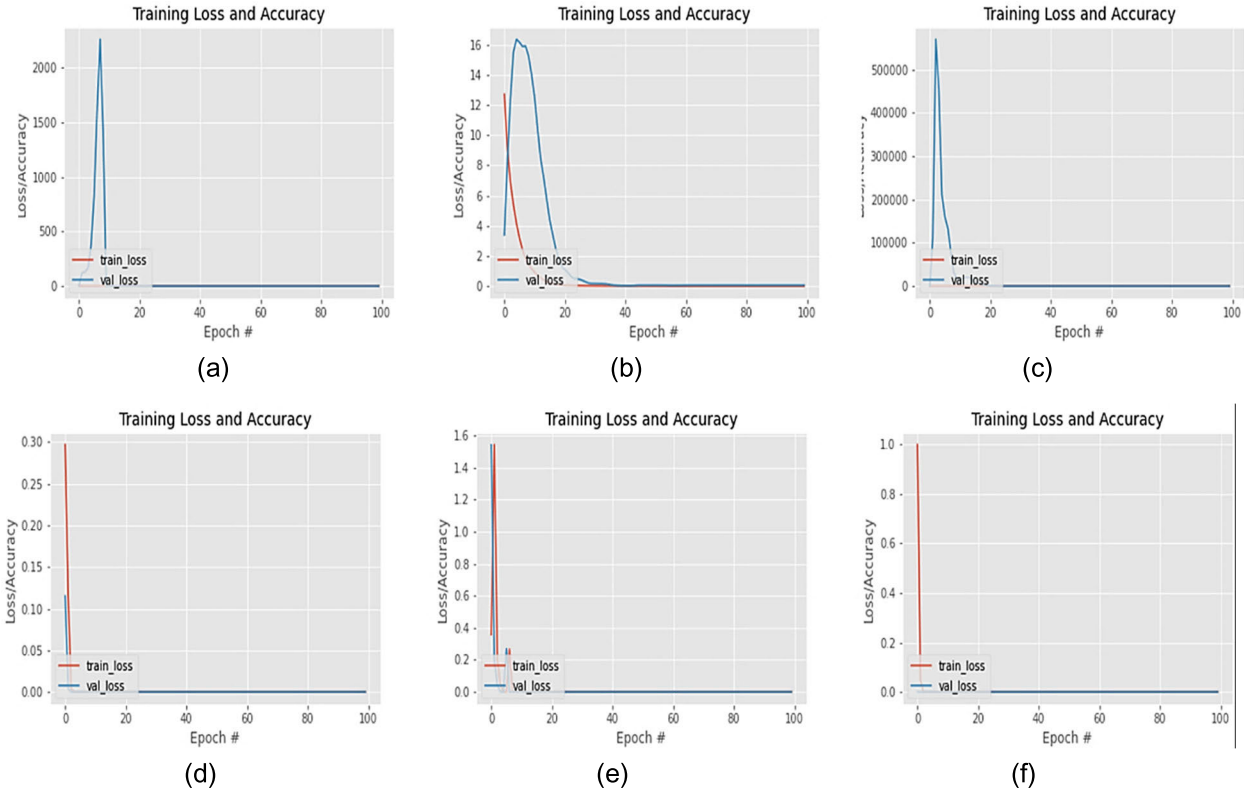


FIGURE 10. Training and validation losses for (a) Inception, (b) Mobilenet, (c) Resnet, (d) VGG16, (e) VGG19, and (f) Xception applied to QuickDraw-Extended dataset.

2) TIME COMPLEXITY

- When applied on the ESRIR dataset, the proposed convolutional autoencoder has lower training and indexing times by about 20.92, 3.755, 16, 19.1, 23.11, and 17.54 times compared to those of InceptionV3, Mobilenet, Resnet50, VGG16, VGG19, and Xception, respectively.
- When applied on the CUFSF dataset, it has lower training and indexing times by about 2.7, 1.5, 5.41, 8, 9.73, and 4.97 times compared to those of InceptionV3, Mobilenet, Resnet50, VGG16, VGG19, and Xception, respectively.
- Similarly, when applied on the LFW dataset, it has lower training and indexing times by about 2.94, 1.75, 6.77, 10.06, 10.86, and 5.452 times compared to those of InceptionV3, Mobilenet, Resnet50, VGG16, VGG19, and Xception, respectively.
- In addition, when applied on the QuickDraw-Extended dataset, it has lower training and indexing times by about 13.35, 9.6, 31.934, 48.214, 95.16, and 26.493 times compared to those of InceptionV3, Mobilenet, Resnet50, VGG16, VGG19, and Xception, respectively.

B. TRAINING COMPLEXITY FOR THE PROPOSED InfoGAN SYSTEM

The suggested InfoGAN contains 145139414 training parameters, as shown in Table 1. When compared to CNN models, the following conclusions may be drawn.

1) TRAINING PARAMETERS

In comparison to InceptionV3, Mobilenet, Resnet50, VGG16, VGG19, and Xception, the suggested InfoGAN has more training parameters by about 6.7, 44, 6.2, 9.9, 7.2, and 6.9 times, respectively.

2) TIME COMPLEXITY

- When applied on ESRIR dataset, it has lower training and indexing times by about 42.7, 7.67, 31.834, 39.16, 47.197, 35.82 times compared to those of InceptionV3, Mobilenet, Resnet50, VGG16, VGG19, and Xception, respectively.
- For CUFSF dataset, it has lower training and indexing times by about 1.45, and 1.78 times compared to those of VGG16, and VGG19, respectively. But it has higher times compared to InceptionV3, Mobilenet, Resnet50, and Xception by about 2, 3.6, 1.01, and 1.1 times, respectively.
- For LFW, it has lower training and indexing times by about 1.5, 2.19, 2.37, and 1.19 times compared to those of Resnet50, VGG16, VGG19, and Xception, respectively. But it has higher times compared to those of InceptionV3 and Mobilenet by about 1.6 and 2.6 times, respectively.
- For QuickDraw-Extended dataset, it has lower training and indexing times by about 3.26, 2.344, 7.78, 11.76, 23.2, and 6.46 times compared to those of InceptionV3,

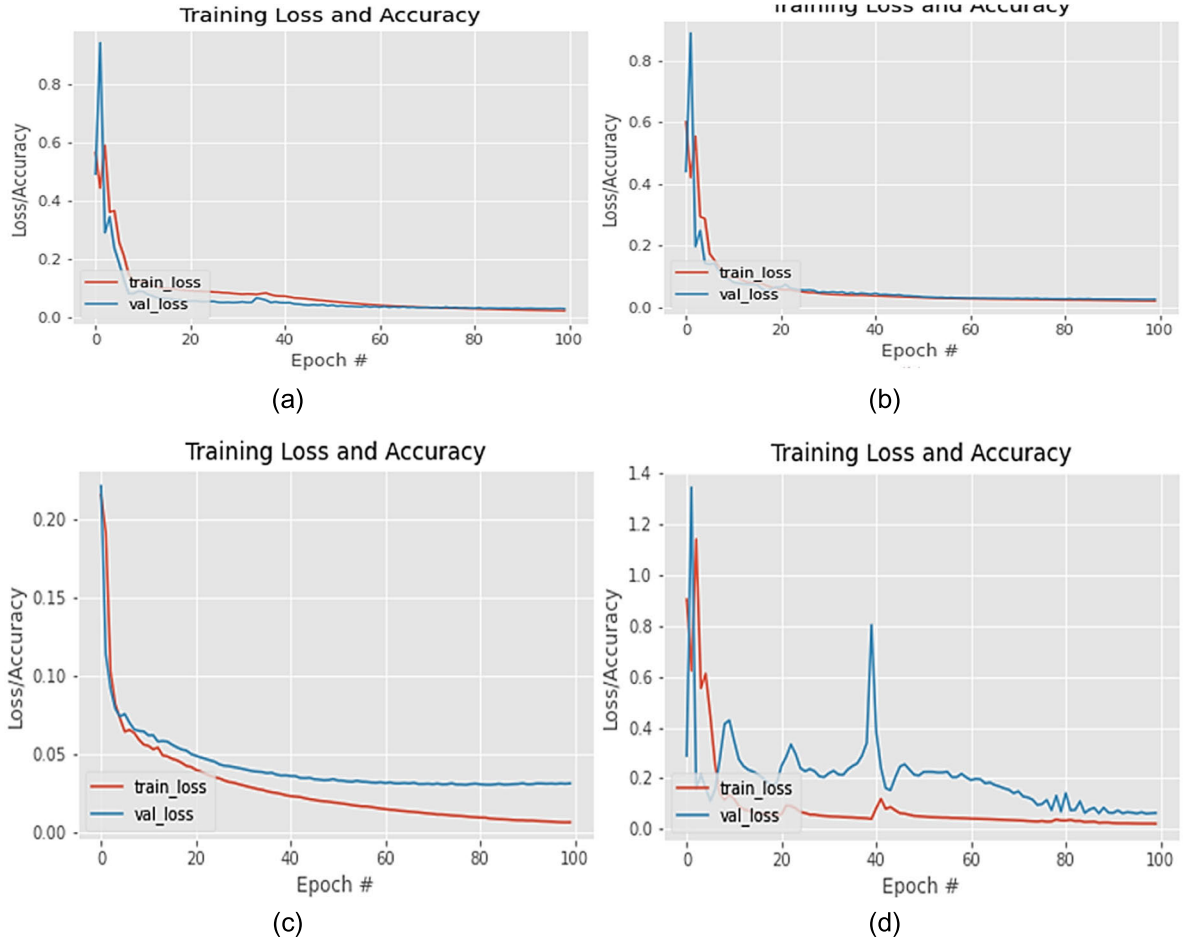


FIGURE 11. Training and validation losses for convolutional autoEncoder applied on (a) ESRIR, (b) CUFSF, (c) LFW, and (d) QuickDraw-Extended datasets.

Mobilenet, Resnet50, VGG16, VGG19, and Xception, respectively.

C. TRAINING COMPLEXITY FOR THE PROPOSED CBIR SYSTEM BASED ON ViT MODEL

As indicated in Table 1, the proposed ViT model has approximately 84,730,372 training parameters. Therefore, in comparison to CNN models, the following notes may be concluded:

1) TRAINING PARAMETERS

Compared to InceptionV3, Mobilenet, Resnet50, VGG16, VGG19, and Xception, the ViT model has higher training parameters of around 4, 26, 4, 5.7, 4.23, and 4.062 times, respectively.

2) TIME COMPLEXITY

- When applied on ESRIR dataset and in comparison to InceptionV3, Resnet50, VGG16, VGG19, and Xception, it reduces training and indexing times by 1.76, 1.31, 1.61, 2, and 1.48 times, respectively. Compared to Mobilenet, on the other hand, it has a 3.17-fold increase.
- For CUFSF dataset, compared to InceptionV3, Mobilenet, Resnet50, VGG16, VGG19, and Xception, the ViT model increases training and indexing times by around 5.14, 9.15, 2.5, 1.72, 1.42, and 2.77 times, respectively.
- When applied to the LFW dataset, the ViT model has higher training and indexing times by about 7.97, 13.4, 3.5, 2.3, 2.16, and 4.3 times compared to those of InceptionV3, Mobilenet, Resnet50, VGG16, VGG19, and Xception, respectively.
- When trained by the QuickDraw-Extended dataset, it has lower training and indexing times by about 1.25, 3, 4.16, 8.92, and 2.5 times compared to those of InceptionV3, Resnet50, VGG16, VGG19, and Xception, respectively. Compared to Mobilenet, it has about 1.11 times increase.

Overall, the significant note is the requirement of powerful processing capabilities, especially with the expansion of dataset size. As a result, cloud computing may be beneficial in such situations. Google cloud offers a variety of cloud services, such as Google storage and processing. Google gives a cloud Tensor Processing Unit (TPU) in a single pod. The TPU and bespoke high-speed network provide over 100 petaflops of performance. In addition, a high-performance Graphics Processing Unit (GPU) is a specialized processor that was created to speed up the rendering of visuals. The GPUs can

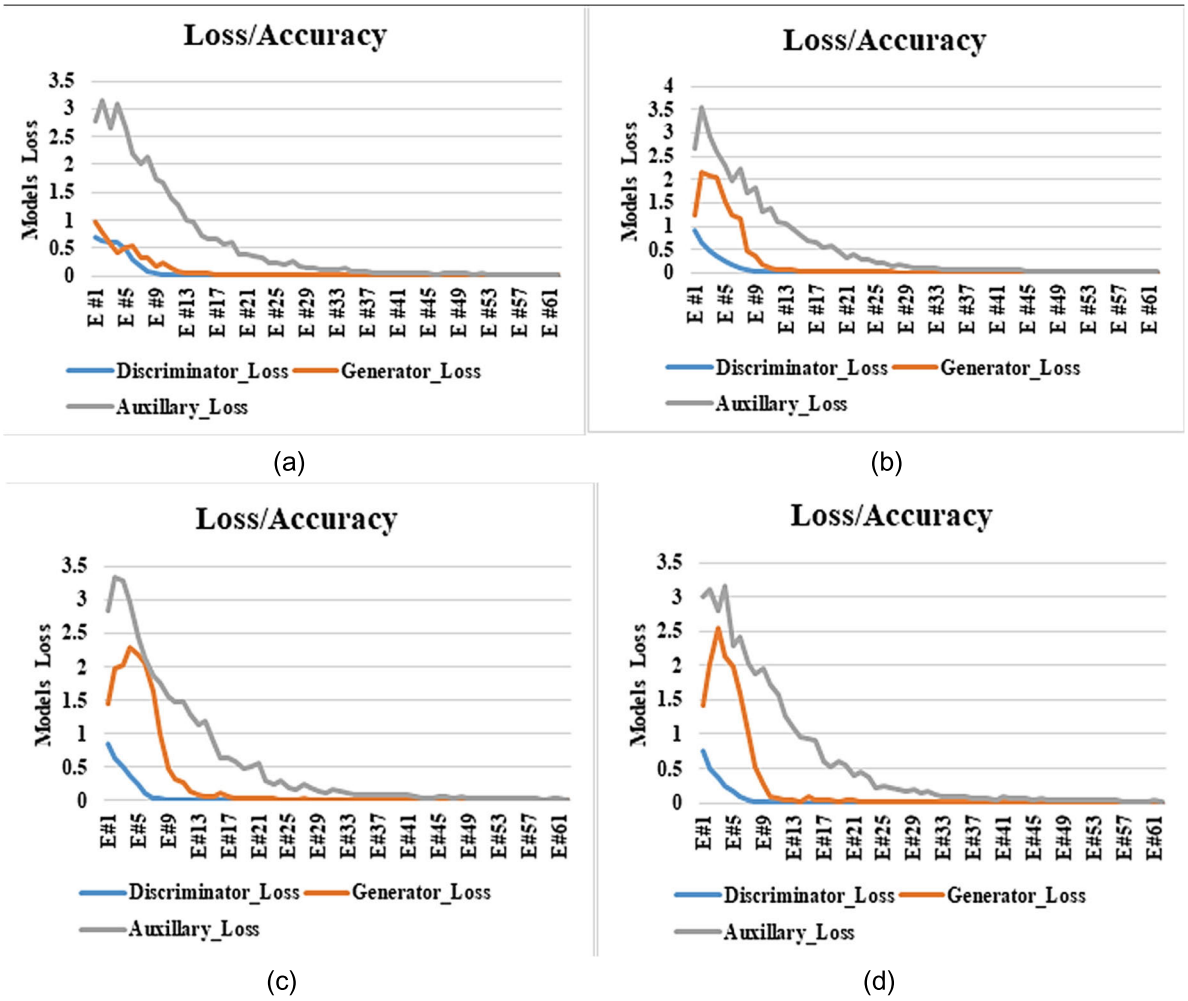


FIGURE 12. Training loss for InfoGAN applied on (a) ESRIR, (b) CUFSF, (c) LFW, and (d) QuickDraw-Extended datasets.

handle a large amount of data at once, making them ideal for machine learning, video editing, and gaming. However, the GPU is used in most situations to speed up processing, so that the amount of training data can be handled. It is simple to use because of the little line of codes.

Furthermore, according to the procedure hierarchy of the three proposed models in Section (VII), each CNN model is used and trained on the ESRIR, CUFSF, LFW, and QuickDraw-Extended datasets. After training, indexing, and spatial distance measurement of query image features, the returned images are all dark and unclear. Hence, in these types of images, CNN models cannot learn to distinguish significant features, which is the key engine for the whole retrieval process.

IX. TEST CASES AND FINDINGS

In this section, performance assessment for three different proposed image-retrieval systems applied on various types of images is involved. Each experimental scenario entails a comparison of different models in terms of various image types matching and retrieval for large-scale datasets. It is important to note that ten query images are chosen randomly from all test instances for each dataset to retrieve their similarities from the dataset using each suggested retrieval system. Because datasets lack comparable images to the query ones, the assessment on CUFSF and LFW datasets will only cover image retrieval not performance metrics computation. After retrieval of images for each randomly-chosen query, True Positive (TP), False Negative (FN), and False Positive (FP) rates are identified in each test case. In other words, the images that are relevant and irrelevant are counted and used to determine the retrieval system performance metrics.

A. 1st TEST CASE ON THE ESRIR Dataset

- Retrieved images of proposed models:

Figures 16, 17, and 18 show samples of the resulting images from the introduced CBIR systems based on convolutional autoencoder, InfoGAN, and ViT retrieval models, respectively. These figures show the retrieved images compared to the query image based on the spatial distance measurement between the features generated by each model. The number of relevant retrieved images using the InfoGAN

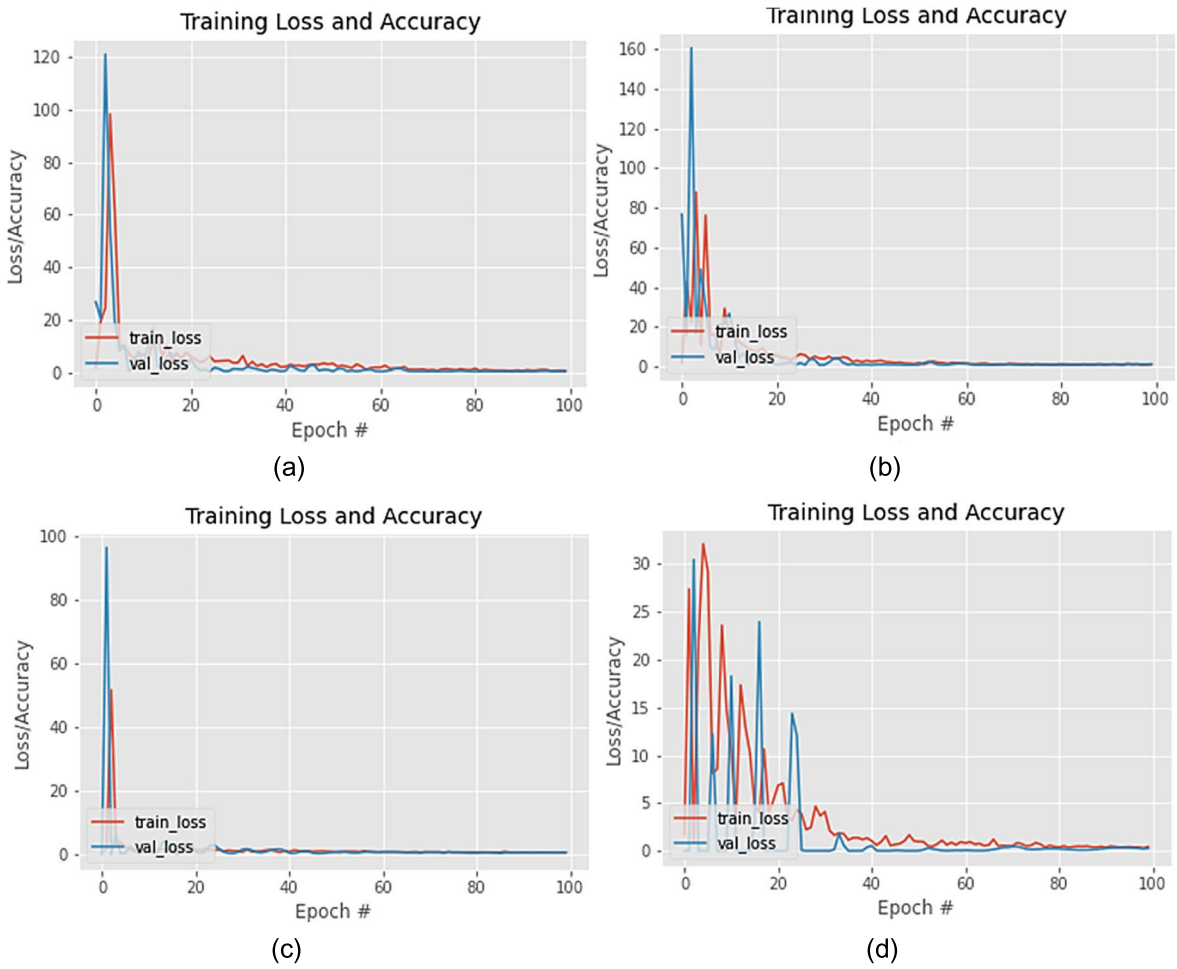


FIGURE 13. Training and validation Losses for ViT applied on (a) ESRIR, (b) CUFSF, (c) LFW, and (d) QuickDraw-Extended datasets.

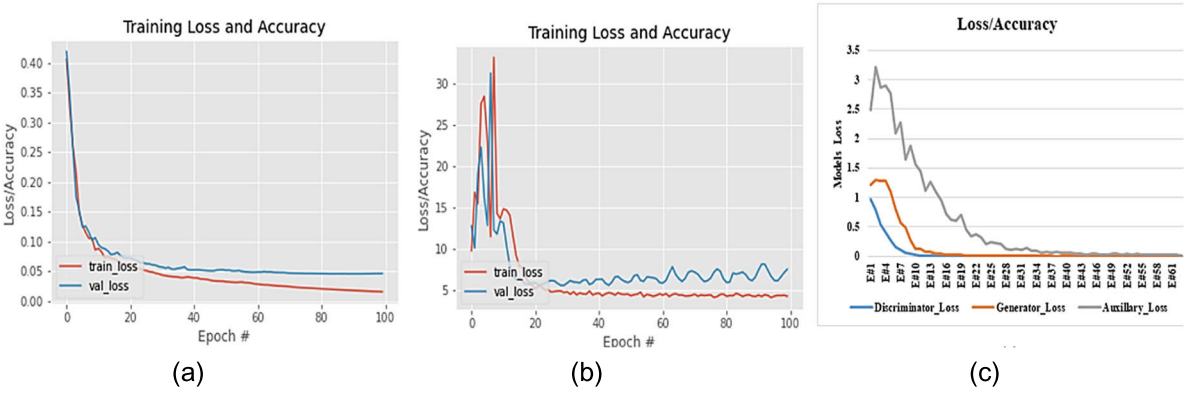


FIGURE 14. Training and validation losses for (a) convolutional autoencoder, (b) ViT, and (c) InfoGAN applied on 256_Object categories.

model is higher than the number of non-relevant ones, as shown in the figures. It is also worth noting that there are few images that are similar, but not like the query image. This implies that when information is retrieved, the highly comparable features are detected and endowed by the InfoGAN model. The ViT is found in the second stage with a slight change in the quantity of relevant and non-related images for the query image comparison. In contrast to InfoGAN, ViT has a lower quantity of enrolled comparable images within the retrieved images. Because the number of relevant retrieved images is limited, the convolutional encoder is in the final stage.

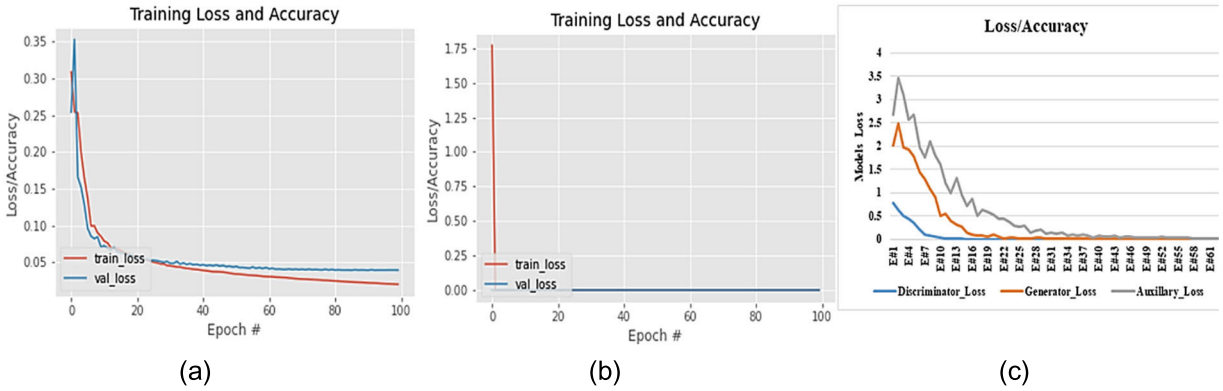


FIGURE 15. Training loss for (a) convolutional autoencoder, (b) ViT, and (c) InfoGAN applied on Flickr Logos 27 dataset.

FIGURE 16. Resulting retrieved images based on a convolutional autoencoder image retrieval system.

FIGURE 17. Resulting retrieved images based on the InfoGAN image retrieval system.

Retrieval performance computation and assessment for the proposed models:

Following the retrieval of images by the three suggested image retrieval systems, Figure 19 displays the computed recall/precision values based on the results. The figure gives the calculated recall/precision for the retrieval process with each query image (i.e., over the 10 query samples) for the ESRIR dataset.

- For the 1st retrieval system based on convolutional autoencoder, Figure 19.a shows the computed recall/precision values after training and retrieval of images. The total computed recall and precision scores are 0.98 and 0.89, respectively.
- For the 2nd retrieval system based on InfoGAN, Figure 19.b shows the obtained recall/precision values after training and retrieval of images. The overall com-

FIGURE 18. Resulting retrieved images based on the ViT image retrieval system.

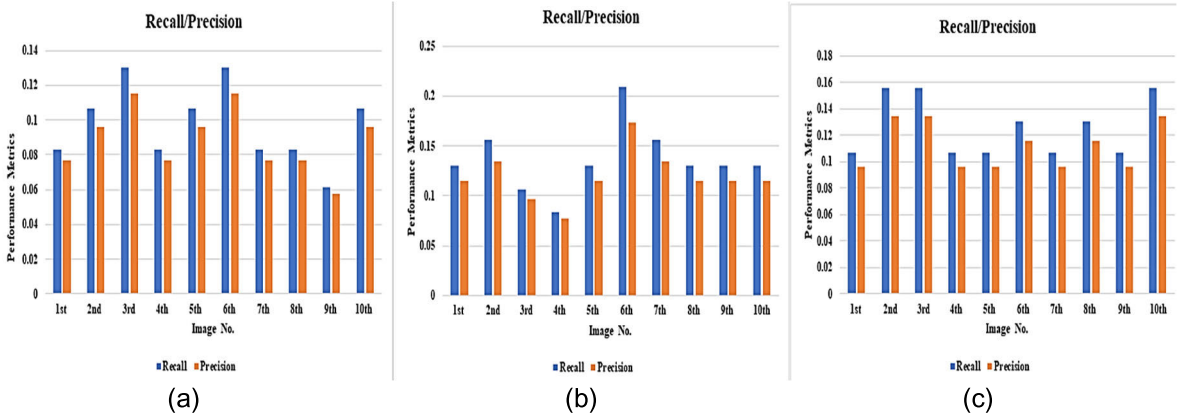


FIGURE 19. Recall/precision on ESRIR dataset with (a) convolutional autoencoder, (b) InfoGAN, and (c) ViT retrieval systems.

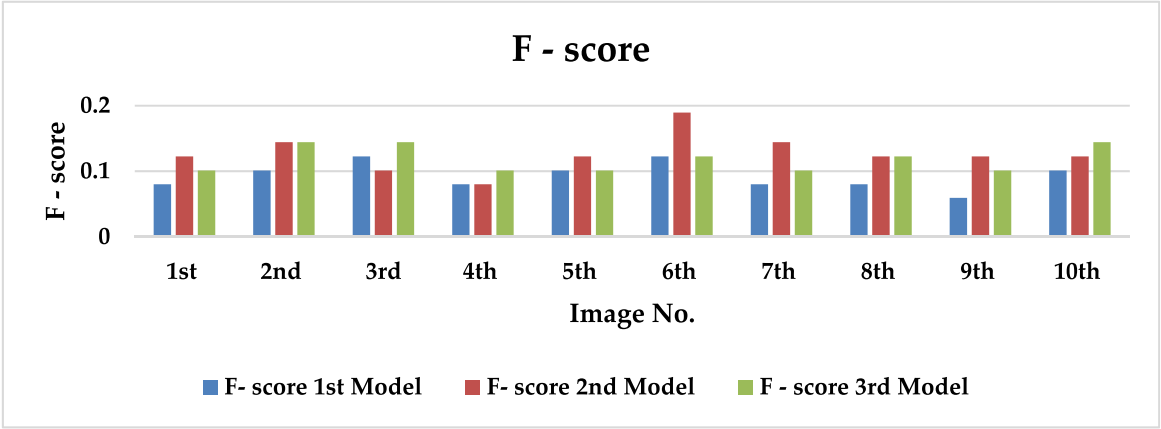


FIGURE 20. F-score on ESRIR dataset with the 1st model (convolutional autoencoder), 2nd model (InfoGAN), and 3rd model (ViT) retrieval systems.

puted recall and precision values are 1.363 and 1.192, respectively.

- For the 3rd retrieval system based on ViT, Figure 19.c shows the computed recall/precision values after training and retrieval of images. The overall recall and precision values are 1.26 and 1.12, respectively.

Figure 20 shows the computed F-score value for each applied query image across the three models. The following notes could be concluded for the ESRIR dataset (over the ten query images):

- The 1st retrieval system based on the convolutional autoencoder has an F-score of 0.93 on all ten images.
- The 2nd retrieval system based on InfoGAN receives an F-score of around 1.272 on all ten images.
- The 3rd retrieval system based on ViT has an F-score of around 1.183 on all ten images.

FIGURE 21. Resulting retrieved images with the convolutional autoencoder image retrieval system on CUFS dataset.

FIGURE 22. Resulting retrieved images with the InfoGAN image retrieval system on CUFSF dataset.

FIGURE 23. Resulting retrieved images with the ViT image retrieval system on CUFSF dataset.

B. 2nd AND 3rd TEST CASES ON THE CUFSF AND LFW DATASES

Retrieved images with the proposed models:

For the CUFSF dataset, Figures 21, 22, and 23 give examples of the recovered images produced by the convolutional autoencoder, InfoGAN, and ViT retrieval systems, respectively. By comparing the figures, it can be seen that the ViT image retrieval system takes the top place, while InfoGAN advances to the next level with a slight difference. The convolutional encoder system is the last in the line. This attributed to the quantity of retrieved facial images with common features like dark hair, dark eyes, etc.

FIGURE 24. Resulting retrieved images with convolutional autoencoder image retrieval system on LFW dataset.

FIGURE 25. Resulting retrieved images with InfoGAN image retrieval system on LFW dataset.

FIGURE 26. Resulting retrieved images with ViT image retrieval system on LFW dataset.

For the LFW dataset, examples of recovered images created by the convolutional autoencoder, InfoGAN, and ViT retrieval systems are shown in Figures 24, 25, and 26, respectively. On such a dataset, learning is performed to

FIGURE 27. Resulting retrieved images with the convolutional autoencoder image retrieval system.

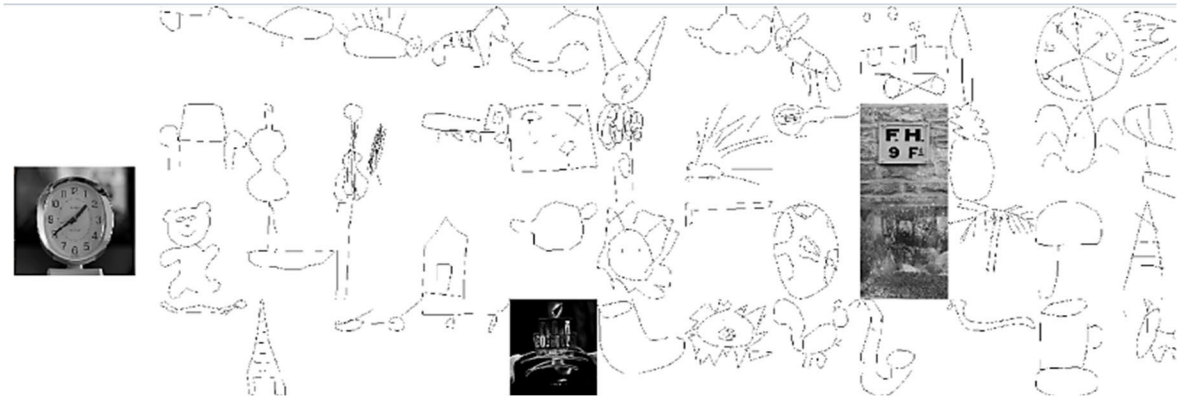


FIGURE 28. Resulting retrieved images with the InfoGAN image retrieval system.

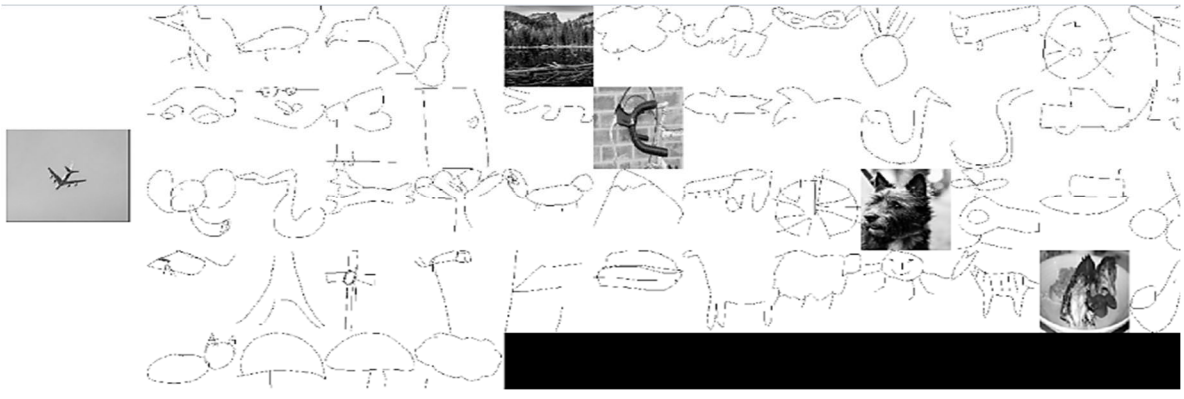


FIGURE 29. Resulting retrieved images wih the ViT image retrieval system.

predict which system is more crucial than the others; yet all three systems return most images with similar features. For example, with the InfoGAN system, the retrieved images are of people with a wide forehead compared to the query images. In comparison to the query images, the retrieved images from the ViT retrieval system are of people with lengthy faces.

C. 4th TEST CASE ON THE QuickDraw-EXTENDED DATASET

- Retrieved images of the proposed models:

Figures 27, 28, and 29 show examples of the recovered images produced by the convolutional autoencoder, InfoGAN, and ViT retrieval systems on the QuickDraw-Extended dataset. On such a dataset, CBIR system based on ViT defeats others as the number of relevant doodled images retrieved

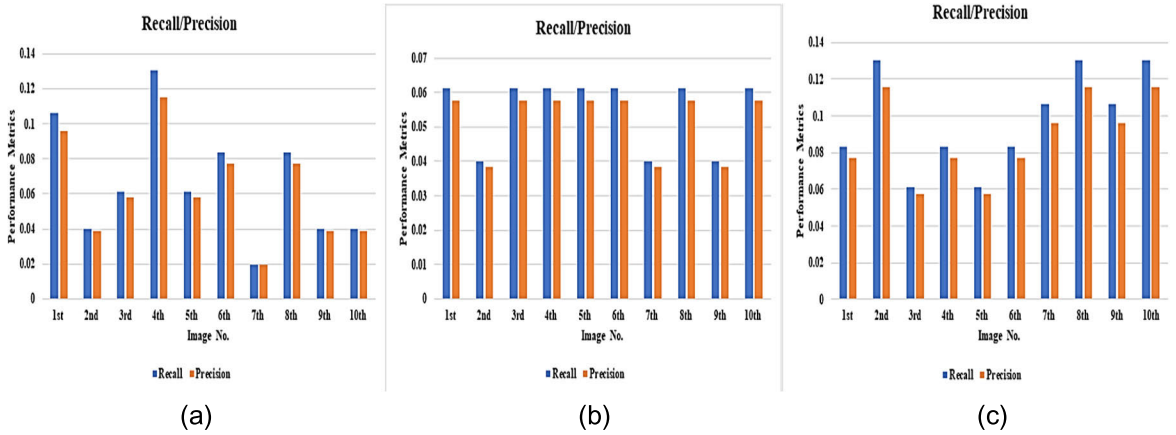


FIGURE 30. Recall/precision on QuickDraw-Extended dataset with (a) convolutional autoencoder, (b) InfoGAN, and (c) ViT retrieval systems.

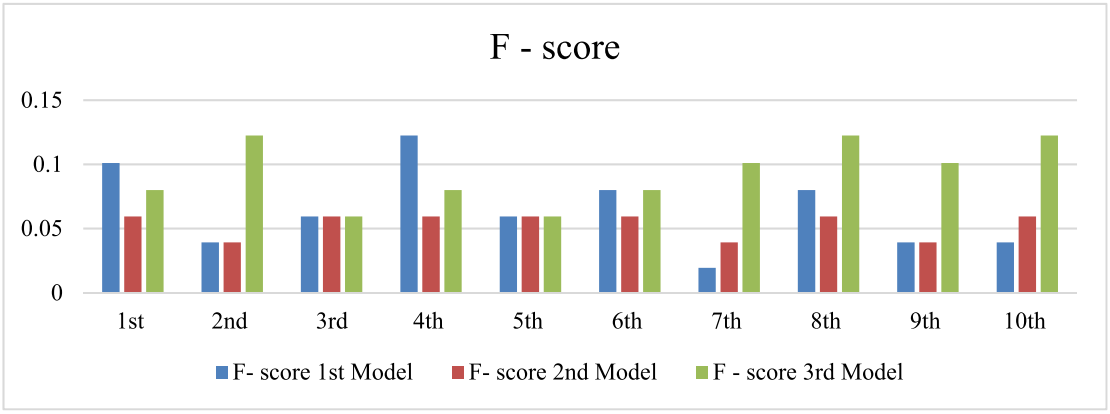


FIGURE 31. F- score on QuickDraw-Extended dataset with the 1st model (convolutional autoencoder), 2nd model (InfoGAN), and 3rd model (ViT) retrieval systems.

FIGURE 32. Resulting retrieved images from 256_Object Categories dataset with the convolutional autoencoder image retrieval system.

is higher than the number of non-relevant ones, as shown in the figures. However, the retrieved images lack some similar images. This is implied by the various retrieved doodlings. Secondly, both convolutional autoencoder and InfoGAN systems are ranked. It is important to highlight that this sort of dataset is complicated to handle, since matching and retrieval are built between real images and doodled images, which are difficult to distinguish by eyesight, much less machine learning.

- Retrieval performance computation and assessment for the proposed models:

Same to same, the computed recall/precision values obtained by the three suggested image retrieval systems after recovering images are shown in Figure 30.

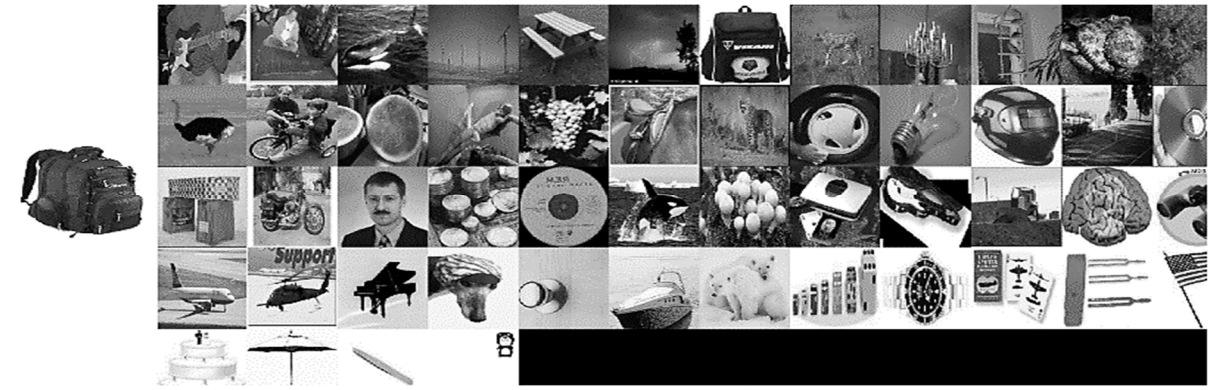


FIGURE 33. Resulting retrieved images from 256_Object Categories dataset with the InfoGAN image retrieval system.

FIGURE 34. Resulting retrieved images from 256_Object Categories datset with the ViT image retrieval system.

FIGURE 35. Resulting retrieved images from Flickr Logos 27 dataset with the convolutional autoencoder image retrieval system.

For the QuickDraw-Extended dataset, Figure 30 shows the computed recall/precision for the retrieved images versus each query image (i.e., over the 10 query samples).

- For the 1st retrieval system based on convolutional autoencoder, Figure 30.a shows the computed recall/precision values. The total recall and precision values are 0.67, and 0.62, respectively.
- For the 2nd retrieval system based on InfoGAN, Figure 30.b displays the calculated recall/precision values. The total recall and precision values are 0.549, and 0.519, respectively.
- For the 3rd retrieval system based on ViT, Figure 30.c displays the calculated recall/precision values. The total recall and precision values are 0.98, and 0.88, respectively.

The computed F-score value for each query image used with the three models is shown in Figure 31. Then, for the QuickDraw-Extended dataset, the following conclusion is obtained:

FIGURE 36. Resulting retrieved images from Flickr Logos 27 with the InfoGAN image retrieval system.

FIGURE 37. Resulting retrieved images from Flickr Logos 27 datset with the ViT image retrieval system.

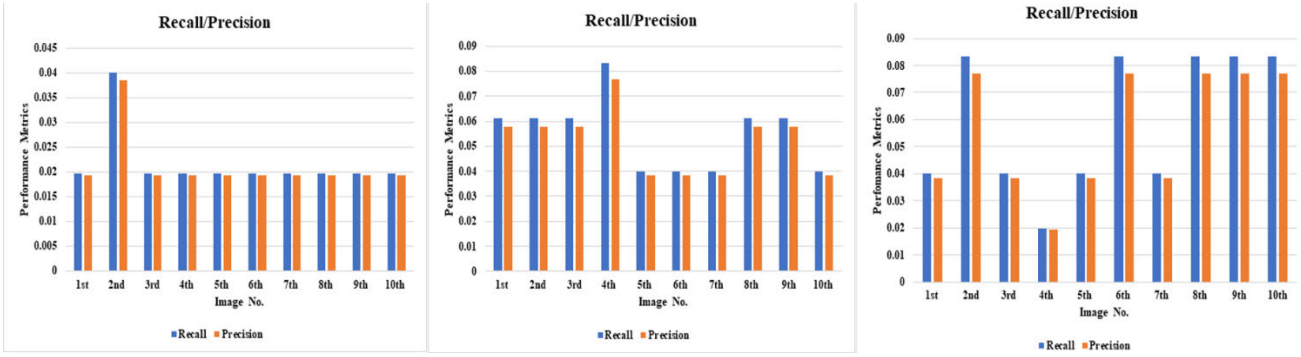


FIGURE 38. Recall/precision on 256_Object Categories dataset with the (a) convolutional autoencoder, (b) InfoGAN, and (c) ViT retrieving systems.

- For the 1st retrieval system based on the the convolutional autoencoder, about 0.64 F-score is achieved with the ten images.
- For the 2nd retrieval system based on InfoGAN, about 0.534 F-score is achieved with the entire ten images.
- For the 3rd retrieval system based on ViT, an F-score of about 0.81 is obtained with the ten images.

D. 5th AND 6th TEST CASES ON THE 256_OBJECT CATEGORIES AND FLICKR LOGOS DATASETS

Retrieved images with the proposed models:

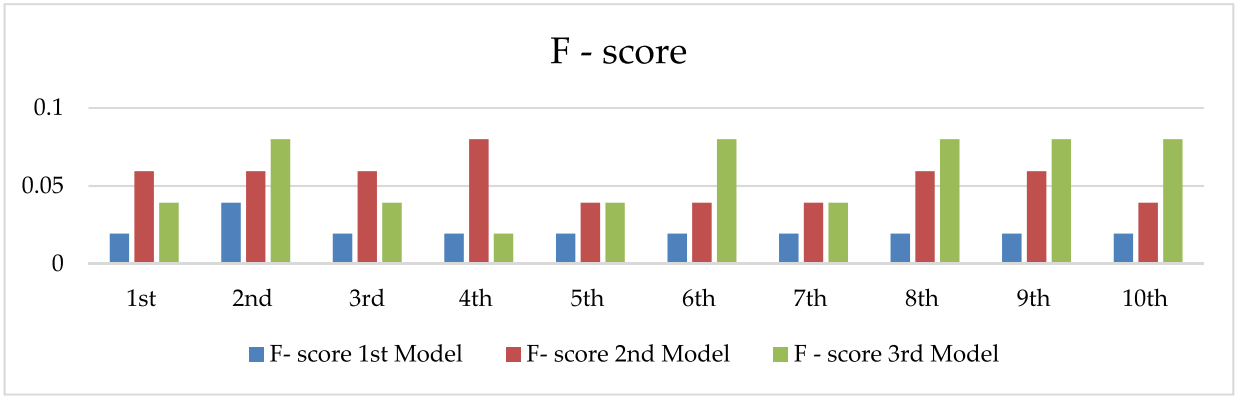


FIGURE 39. F- score on 256_Object Categories dataset with the 1st model (convolutional autoencoder), 2nd model (InfoGAN), and 3rd model (ViT) retrieval systems.

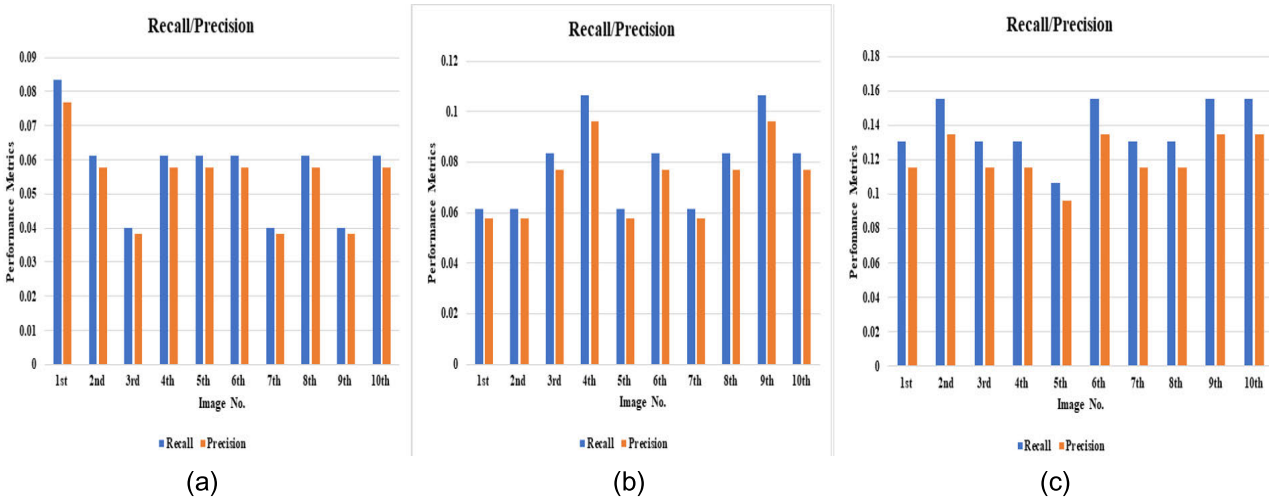


FIGURE 40. Recall/precision on Flickr Logos 27 dataset with the (a) convolutional autoencoder, (b) InfoGAN, and (c) ViT retrieval systems.

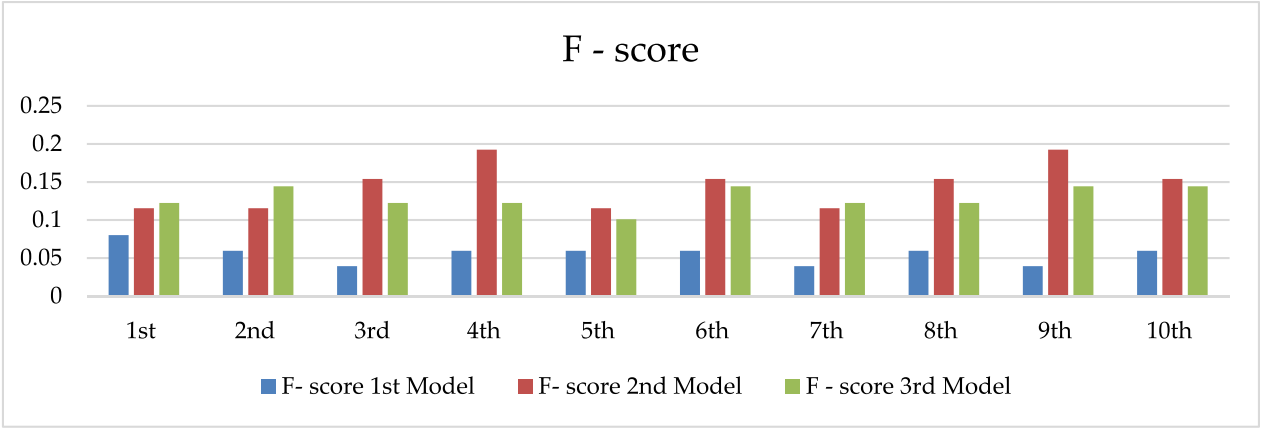


FIGURE 41. F- score on Flickr Logos 27 dataset with the 1st model (convolutional autoencoder), 2nd model (InfoGAN), and 3rd model (ViT) retrieval systems.

- For the 256_Object Categories dataset, Figures 32, 33, and 34 show examples of recovered images created by the convolutional autoencoder, InfoGAN, and ViT retrieval systems. The figures show that ViT and InfoGAN are very close in the query number of relevant and non-relevant images. At the same time, the convolutional

TABLE 1. Training parameters for all tested models.

Model		Trainable parameters	Training datasets					
			1st set (ESRIR)	2nd set (CUFSF)	3rd set (LFW)	4th set (QuickDraw-Extended)	5th 256 Categories	6th Flickr Logos 27
			Training and Indexing Time (s)	Training and Indexing Time (s)	Training and Indexing Time (s)	Training and Indexing Time (s)	Training and Indexing Time (s)	Training and Indexing Time (s)
Convolutional autoencoder	Encoder	11,005,184	238.403	190.3302	238.396	267.933	197.76	207.17
	Decoder	11,025,153						
Vision Transformer		84,730,372	2835.57	2621.987	5590.91	2858.52	3134.04	2615.593
Infogans	Generative Model	53,750,849	116.766	1043.44	1092.746	1098.548	1029.784	1127.76
	Auxiliary Model	45,518,666						
	Discriminative Model	1,008,065						
	GAN Model	98,612,683						
InceptionV3		21,802,784	4987.088	510.2096	701.377	3577.309	Not trained	
Mobilenet		3,228,864	895.189	286.486	418.029	2574.655		
Resnet50		23,587,712	3717.113	1029.9304	1612.862	8556.063		
VGG-16		14,714,688	4572.729	1523.308737	2398.458	12918.136		
VGG-19		20,024,384	5510.995	1851.3015	2589.992	25495.5596		
Xception		20,861,480	4182.579	945.2174	1299.666	7098.413		

autoencoder comes at the end of the line. For simplicity, retrieval on such a dataset is assessed over the random selection for the query images, not for each category of the 256.

- For the Flickr Logos 27 dataset, Figures 35, 36, and 37 show examples of the recovered images by the convolutional autoencoder, InfoGAN, and ViT retrieval systems. The figures show that ViT surpasses other systems on these types of images.

Retrieval performance computation and assessment for the proposed models:

- For the 256 Object Categories dataset (for ten query images), Figure 38 shows the computed values attained by the three image retrieval systems on the 10 query samples. Figure 39 shows the computed F-score value for each used query image with the three models.

Computed recall/precision values:

– For the convolutional autoencoder retrieval system, the total recall and precision values are 0.22 and 0.21, respectively.
– For the InfoGAN retrieval system, the total recall and precision values are 0.55 and 0.52, respectively.
– For the ViT retrieval system, the total recall and precision values are 0.597 and 0.558, respectively.

F-score value:

– The convolutional autoencoder retrieval system achieves an F-score of about 0.22 on the entire ten images.
– The InfoGAN retrieval system achieves an F-score of about 0.533 on the entire ten images.
– The ViT retrieval system achieves an F-score of about 0.497 on the entire ten images.

- On the Flickr Logos 27 dataset (for ten query images), similarly, Figure 40 shows the calculated recall/precision over the entire 10 query images. Figure 41 displays the computed F-score values for the same samples.

Computed recall/precision values:

– By applying the convolutional autoencoder retrieval system, the total recall and precision values are 0.57 and 0.54, respectively.
– For the InfoGAN case, the total recall and precision values are 0.79 and 0.73, respectively. Besides, it achieves an F-score of about 1.462 for all 10 images.
– For the ViT retrieval system, the total recall and precision values are 1.381 and 1.21, respectively. Besides, it achieves an F-score of about 1.147 over all ten images.

F-score value:

– The convolutional autoencoder retrieval system achieves an F-score of about 0.56 over the ten images.

– The InfoGAN retrieval system achieves an F-score of about 1.462 for all 10 images.
– The ViT retrieval system achieves an F-score of about 1.147 over all 10 images.

X. RESULT DISCUSSION

The F-score is defined as the harmonic mean of precision and recall for image retrieval system performance assessment. Higher F-scores are, in fact, necessary for improved performance. These scores can vary from 0 to 1, with 1 indicating a model that flawlessly classifies each observation into the correct class and 0 indicating a model that cannot classify any observation into the correct class. It is a better metric than accuracy. Both of those metrics use class predictions as input by comparing such scores to correctness. On the other hand, the F-score balances precision and recall for the positive class, whereas accuracy looks at properly identified positive and negative observations. As a result, F-score may be the greatest discriminating statistic tool for making appropriate judgments on retrieval system performance. Furthermore, the recall/precision should allow a strong assessment for such retrieval system performance. According to the acquired findings:

– The InfoGAN retrieval system outperforms other corresponding systems in obtaining similar images when comparing facial sketches to real images, as shown from the results on the ESRIR dataset. Therefore, the second rank goes to the ViT retrieval system and the last to the convolutional encoder. However, there is a slight difference between the CBIR systems based on InfoGAN and ViT models.
– On the QuickDraw-Extended dataset, it is found that the ViT retrieval system outperforms the other systems. InfoGAN and convolutional autoencoder systems come after with slight differences.
– The ViT and InfoGAN retrieval systems are close to each other in performance on the 256_Object Categories dataset. However, there is a slight difference between the ViT and InfoGAN systems. Convolutional autoencoder comes in the second rank.
– In the Flickr Logos 27 case, the ViT retrieval system outperforms the other systems. InfoGAN comes in the second place with a significant performance gap. Again, the convolutional autoencoder is at the bottom of the list.

XI. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

This article might serve as the basis for a broad range of applications that employ features to classify and retrieve objects in images. According to the findings, the key engine of the entire process is the capacity of extraction methods to adequately characterize image content with suitable feature descriptors to increase performance accuracy. Aiming for model generality, experiments demonstrate various retrieval process hurdles that significantly influence the retrieval accuracy for faces and objects in different images. The ESRIR dataset, which includes 53,000 face sketches and 53,000 real facial images, has been presented to the community in order to increase the scale of facial sketched-real image retrieval. Besides, three different image retrieval systems have been proposed in this paper based on convolutional autoencoder, InfoGAN, and ViT. The proposed models have been trained with six different datasets, including the introduced ESRIR dataset. According to the findings, InfoGAN and ViT retrieval systems are more successful in differentiating freehand facial sketch drawings and objects on CUFSF, and the 256_Object Categories datasets. Besides, their outstanding performance on ESRIR dataset is independent of the applied augmentation and visual scene transformations. On the ESRIR dataset, the ViT system achieves about a 1.183 F-score value, whereas the InfoGAN system reaches around a 1.272 F-score. The ViT retrieval system outperforms the other ones on the too-challenging QuickDraw-Extended dataset. The ViT system successfully retrieved images for the 10 query images from the QuickDraw-Extended dataset with an F-score of around 0.81.

The use of other distance metrics in place of the Euclidian distance utilized in this article for convolutional autoencoder, InfoGAN, and ViT instances, is one of the additions to this paper. As an alternative, combined feature extraction algorithms, such as capsule networks, might be employed in conjunction with the ones recommended in this article to investigate and benefit from their efficacy and results. Additionally, several artificial intelligence methods might be investigated on different other datasets.

ACKNOWLEDGMENT

The authors would like to acknowledge the Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R66), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

REFERENCES

[1] N. F. Soliman, M. Khalil, A. D. Algarni, S. Ismail, R. Marzouk, and W. El-Shafai, "Efficient HEVC steganography approach based on audio compression and encryption in QFFT domain for secure multimedia communication," *Multimedia Tools Appl.*, vol. 80, no. 3, pp. 4789–4823, 2020.
[2] W. El-Shafai, "Joint adaptive pre-processing resilience and post-processing concealment schemes for 3D video transmission," *3D Res.*, vol. 6, no. 1, pp. 1–13, Mar. 2015.
[3] K. M. Abdelwahab, S. M. A. El-Atty, W. El-Shafai, S. El-Rabaie, and F. E. A. El-Samie, "Efficient SVD-based audio watermarking technique in FRT domain," *Multimedia Tools Appl.*, vol. 79, nos. 9–10, pp. 5617–5648, Mar. 2020.
[4] A. D. Algarni, G. El Banby, S. Ismail, W. El-Shafai, F. E. A. El-Samie, and N. F. Soliman, "Discrete transforms and matrix rotation based cancelable face and fingerprint recognition for biometric security applications," *Entropy*, vol. 22, no. 12, p. 1361, Nov. 2020.
[5] N. A. El-Hag, A. Sedik, W. El-Shafai, H. M. El-Hoseny, A. A. Khalaf, A. S. El-Fishawy, W. Al-Nuaimy, F. E. A. El-Samie, and G. M. El-Banby, "Classification of retinal images based on convolutional neural network," *Microsc. Res. Technique*, vol. 84, no. 3, pp. 394–414, 2021.
[6] W. El-Shafai, S. El-Rabaie, M. El-Halawany, and F. E. A. El-Samie, "Enhancement of wireless 3D video communication using color-plus-depth error restoration algorithms and Bayesian Kalman filtering," *Wireless Pers. Commun.*, vol. 97, no. 1, pp. 245–268, Nov. 2017.
[7] G. E. Trahey, K. R. Nightingale, R. W. Nightingale, and M. Palmeri, "Method and apparatus for the identification and characterization of regions of altered stiffness," U.S. Patent 6 951 544 Oct. 4, 2005.

[8] J. L. Schonberger, H. Hardmeier, T. Sattler, and M. Pollefeys, "Comparative evaluation of hand-crafted and learned local features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017.

[9] W. El-Shafai, S. El-Rabaie, M. M. El-Halawany, and F. E. A. El-Samie, "Recursive Bayesian filtering-based error concealment scheme for 3D video communication over severely lossy wireless channels," *Circuits, Syst., Signal Process.*, vol. 37, no. 11, pp. 4810–4841, Nov. 2018.

[10] M. K. Patrick, A. F. Adekoya, A. A. Mighty, and B. Y. Edward, "Capsule networks—A survey," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 1, pp. 1295–1310, 2019.

[11] W. El-Shafai, E. M. El-Rabaie, M. M. El-Halawany, and F. E. A. El-Samie, "Proposed adaptive joint error-resilience concealment algorithms for efficient colour-plus-depth 3D video transmission," *IET Image Process.*, vol. 12, no. 6, pp. 967–984, Jun. 2018.

[12] W. El-Shafai, S. El-Rabaie, M. M. El-Halawany, and F. E. Abd El-Samie, "Performance evaluation of enhanced error correction algorithms for efficient wireless 3D video communication systems," *Int. J. Commun. Syst.*, vol. 31, no. 1, p. e3396, Jan. 2018.

[13] S. Torsten, Z. Qunjie, P. Marc, and L.-T. Laura, "Understanding the Limitations of CNN-based absolute camera pose regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3302–3312.

[14] M. Maggipinto, C. Masiero, A. Beghi, and G. A. Susto, "A convolutional autoencoder approach for feature extraction in virtual metrology," *Proc. Manuf.*, vol. 17, pp. 126–133, Jan. 2018, doi: 10.1016/j.promfg.2018.10.023.

[15] A. Creswell and A. A. Bharath, "Adversarial training for sketch retrieval," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 798–809.

[16] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023, doi: 10.1109/TPAMI.2022.3152247.

[17] A. El-Nouby, N. Neverova, I. Laptev, and H. Jégou, "Training vision transformers for image retrieval," 2021, *arXiv:2102.05644*.

[18] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, "A descriptor for large scale image retrieval based on sketched feature lines," in *Proc. 6th Eurographics Symp. Sketch-Based Int. Model.*, Aug. 2009, pp. 29–36.

[19] A. Chalechale, G. Naghdy, and A. Mertins, "Edge image descriptionusing angular radial partitioning," *IEE Proc.-Vis., Image Signal Process.*, vol. 151, no. 2, pp. 93–101, Apr. 2004.

[20] Y. Zhang, X. Qian, X. Tan, J. Han, and Y. Tang, "Sketch-based image retrieval by salient contour reinforcement," *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1604–1615, Aug. 2016, doi: 10.1109/TMM.2016.2568138.

[21] Y. Cao, C. Wang, L. Zhang, and L. Zhang, "Edgel index for large-scale sketch-based image search," in *Proc. CVPR*, Jun. 2011, pp. 761–768.

[22] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, "Sketch-based image retrieval: Benchmark and bag-of-features descriptors," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 11, pp. 1624–1636, Nov. 2011.

[23] S. Deniziak and T. Michno, "New content-based image retrieval database structure using query by approximate shapes," in *Proc. Commun. Papers Federated Conf. Comput. Sci. Inf. Syst.*, vol. 13, M. Ganzha, L. Maciaszek, and M. Paprzycki Eds. Sep. 2017, pp. 177–182.

[24] Y. Liu, D. Yu, X. Chen, Z. Li, and J. Fan, "TOP-SIFT: The selected SIFT descriptor based on dictionary learning," *Vis. Comput.*, vol. 35, no. 5, pp. 667–677, May 2019.

[25] N. Sarafianos, X. Xu, and A. I. Kakadiaris, "Adversarial representation learning for text-to-image matching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5814–5824.

[26] B. Ay, G. Aydin, Z. Koyun, and M. Demir, "A visual similarity recommendation system using generative adversarial networks," in *Proc. Int. Conf. Deep Learn. Mach. Learn. Emerg. Appl. (Deep-ML)*, Aug. 2019, pp. 44–48, doi: 10.1109/deep-ml.2019.00017.

[27] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Z. Jiang, Q. Hou, and J. Feng, "DeepViT: Towards deeper vision transformer," 2021, *arXiv:2103.11886*.

[28] S. H. Lee, S. Lee, and B. C. Song, "Vision transformer for small-size datasets," 2021, *arXiv:2112.13492*.

[29] S. U. Rehman, S. Tu, M. Waqas, Y. Huang, O. U. Rehman, B. Ahmad, and S. Ahmad, "Unsupervised pre-trained filter learning approach for efficient convolution neural network," *Neurocomputing*, vol. 365, pp. 171–190, Nov. 2019, doi: 10.1016/j.neucom.2019.06.084.

[30] S. U. Rehman, S. Tu, Y. Huang, and G. Liu, "CSFL: A novel unsupervised convolution neural network approach for visual pattern classification," *AI Commun.*, vol. 30, no. 5, pp. 311–324, Aug. 2017, doi: 10.3233/AIC-170739.

[31] S. Rehman, S. Tu, O. Rehman, Y. Huang, C. Magurawalage, and C.-C. Chang, "Optimization of CNN through novel training strategy for visual classification problems," *Entropy*, vol. 20, no. 4, p. 290, Apr. 2018, doi: 10.3390/e20040290.

[32] S. U. Rehman, S. Tu, Y. Huang, and Z. Yang, "Face recognition: A novel un-supervised convolutional neural network method," in *Proc. IEEE Int. Conf. Online Anal. Comput. Sci. (ICOACS)*, May 2016, pp. 139–144, doi: 10.1109/ICOACS.2016.7563066.

[33] N. Keisham and A. Neelima, "Efficient content-based image retrieval using deep search and rescue algorithm," *Soft Comput.*, vol. 26, no. 4, pp. 1597–1616, Feb. 2022, doi: 10.1007/s00500-021-06660-x.

[34] K. N. Sukhia, S. S. Ali, M. M. Riaz, A. Ghafoor, and B. Amin, "Content-based image retrieval using angles across scales," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2021.3131340.

[35] *Deep-Learning-Based Machine Understanding of Sketches: Recognizing and Generating Sketches With Deep Neural Networks*. Accessed: May 2022. [Online]. Available: https://www2.eecs.berkeley.edu/Pubs/TechRpts/2020/EECS-2020-13.pdf

[36] A. R. Sharma and P. R. Devale, "Face photo-sketch synthesis and recognition," *Int. J. Appl. Inf. Syst.*, vol. 1, no. 6, pp. 46–52, Feb. 2012.

[37] S. U. Rehman, S. Tu, Y. Huang, and O. U. Rehman, "A benchmark dataset and learning high-level semantic embeddings of multimedia for cross-media retrieval," *IEEE Access*, vol. 6, pp. 67176–67188, 2018, doi: 10.1109/ACCESS.2018.2878868.

[38] T. D. Gedeon and D. Harris, "Progressive image compression," in *Proc. Int. Joint Conf.*, vol. 4, Jun. 1992, pp. 403–407.

[39] B. Dillon, T. Plehn, C. Sauer, and P. Sorrenson, "Better latent spaces for better autoencoders," *SciPost Phys.*, vol. 11, no. 3, p.061, 2021, doi: 10.21468/SciPostPhys.11.3.061.

[40] A. Radoi, "Convolutional autoencoder-based image reconstruction for unsupervised multimodal change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2021, pp. 4372–4375, doi: 10.1109/IGARSS47720.2021.9553400.

[41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[42] S. Gkelios, Y. Boutalis, and S. A. Chatzichristofis, "Investigating the vision transformer model for image retrieval tasks," in *Proc. 17th Int. Conf. Distrib. Comput. Sensor Syst. (DCOSS)*, 2021, pp. 367–373.

[43] G. Salton, *The SMART Retrieval System, Experiments in Automatic Document Processing*. Upper Saddle River, NJ, USA: Prentice-Hall.

[44] C.J. Van Rijsbergen, *Evaluation, Information Retrieval*. Upper Saddle River, NJ, USA: Prentice-Hall, 1979, ch. 7, pp. 112–123.

[45] W.-S. Hwang, J. J. Weng, M. Fang, and J. Qian, "A fast image retrieval algorithm with automatically extracted discriminant features," in *Proc. IEEE Workshop Content-Based Access Image Video Libraries (CBAIVL)*, Fort Collins, CO, USA, Jun. 1999, pp. 8–12.

[46] S. Müller and G. Rigoll, "Improved stochastic modeling of shapes for content-based image retrieval," in *Proc. IEEE Workshop Content-Based Access Image Video Libraries (CBAIVL)*, Fort Collins, CO, USA, Jun. 1999, pp. 23–27.

[47] B. Ozer, W. Wolf, and A. N. Akansui, "A graph based object description for information retrieval in digital image and video libraries," in *Proc. IEEE Workshop Content-Based Access Image Video Libraries (CBAIVL)*, Fort Collins, CO, USA, Jun. 1999, pp. 79–83.

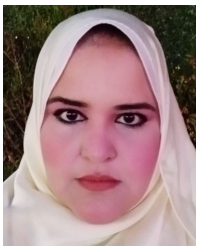
[48] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Proc. CVPR*, Jun. 2011, pp. 513–520.

[49] *Labeled Faces in the Wild Home*. Accessed: Jan. 2022. [Online]. Available: http://vis-www.cs.umass.edu/lfw/

[50] *Doodle to Search: Practical Zero-Shot Sketch-Based Image Retrieval*. Accessed: Feb. 2022. [Online]. Available: https://sounakdey.github.io/doodle2search.github.io/

[51] *Caltech 256 Image Dataset*. Accessed: Feb. 2022. [Online]. Available: https://www.kaggle.com/jessicali9530/caltech256

[52] Y. Kalantidis, L. G. Pueyo, M. Trevisiol, R. Van Zwol, and Y. Avrithis, "Scalable triangulation-based logo recognition," in *Proc. 1st ACM Int. Conf. Multimedia Retr.*, Trento, Italy, Apr. 2011, pp. 1–7.



EMAN S. SABRY was born in Cairo, Egypt. She received the B.Sc. degree (Hons.) in electronics and electrical communication engineering from the Higher Institute of Engineering, El Shorouk Academy, and the M.Sc. degree from the Faculty of Engineering, Arab Academy for Science, Technology, and Maritime Transport, in 2016. She is currently working as a Teaching Assistant at the Higher Institute of Engineering, El Shorouk Academy. She has competence utilizing Python to create current deep learning models across a variety of phases and applications. She also has experience in communication systems, microwave, antenna, optical signal processing, electronics, image and video compression, cloud computing, multimedia processing, and other areas. She also has expertise in developing and processing IP networks, such as IPTV, using OPNET and various network simulators. Her current research interests include artificial intelligence, deep learning, generative adversarial networks (GAN), vision transformer, capsule networks, data analysis, image processing, and retrieval.

SALAH S. ELAGOOZ received the B.Sc. and M.Sc. form MTC, Cairo, Egypt, in 1981 and 1987, respectively, and the Ph.D. degree in efficient communication systems from GWU, Washington, DC, USA, in 1993. He is currently the Head of Communications and Computer Engineering Department, El-Shorouk High Engineering Institute. His areas of interest are mobile and satellite communication systems, channel coding, and encryption.

FATHI E. ABD EL-SAMIE received the B.Sc. (Hons.), M.Sc., and Ph.D. degrees from the Faculty of Electronic Engineering, Menoufia University, Menouf, Egypt, in 1998, 2001, and 2005, respectively. He joined the Teaching Staff of the Department of Electronics and Electrical Communications, Faculty of Electronic Engineering, Menoufia University, in 2005. His current research interests include image enhancement, image restoration, image interpolation, super resolution reconstruction of images, data hiding, multimedia communications, medical image processing, optical signal processing, and digital communications. He has received the Most Cited Paper Award from *Digital Signal Processing* journal, in 2008.

WALID EL-SHAFAI was born in Alexandria, Egypt. He received the B.Sc. degree (Hons.) in electronics and electrical communication engineering from the Faculty of Electronic Engineering (FEE), Menoufia University, Menouf, Egypt, in 2008, the M.Sc. degree from the Egypt–Japan University of Science and Technology (E-JUST), in 2012, and the Ph.D. degree from the Faculty of Electronic Engineering, Menoufia University, in 2019. Since January 2021, he has been a Postdoctoral Research Fellow at the Security Engineering Laboratory (SEL), Prince Sultan University (PSU), Riyadh, Saudi Arabia. He is currently working as a Lecturer and an Assistant Professor with the Electronics and Communication Engineering (ECE) Department, FEE, Menoufia University. His research interests include wireless mobile and multimedia communications systems, image and video signal processing, efficient 2-D video/3-D multiview video coding, multi-view video plus depth coding, 3-D multi-view video coding and transmission, quality of service and experience, digital communication techniques, cognitive radio networks, adaptive filters design, 3-D video watermarking, steganography, and encryption, error resilience and concealment algorithms for H.264/AVC, H.264/MVC, and H.265/HEVC video codecs standards, cognitive cryptography, medical image processing, speech processing, security algorithms, software-defined networks, the Internet of Things, medical diagnoses applications, FPGA implementations for signal processing algorithms and communication systems, cancellable biometrics and pattern recognition, image and video magnification, artificial intelligence for signal processing algorithms and communication systems, modulation identification and classification, image and video super-resolution and denoising, cybersecurity applications, malware and ransomware detection and analysis, deep learning in signal processing, and communication systems applications. He has several publications in the above research areas in several reputable international and local journals and conferences. He serves as a reviewer for several international journals.

NIRMEEN A. EL-BAHNASAWY received the B.S. degree in electronic engineering, in 1998, and the M.Sc. and Ph.D. degrees in computer science and engineering from Menoufia University, in 2003 and 2013, respectively. She was appointed as an Associate Professor at Menoufia University, in 2019. She has deep experience in dealing with electronics H/W kits, different software tools, and different programming languages. She did and supervised different H/W and S/W implementation projects. Her research interests include distributed computing, grid computing, IoT, artificial intelligence, fog computing, and cloud computing.

GHADA M. EL-BANBY received the M.Sc. and Ph.D. degrees in automatic control engineering from Menoufia University, Egypt, in 2006 and 2012, respectively. She is currently working as an Associate Professor at the Department of Industrial Electronics and Control Engineering, Faculty of Electronic Engineering, Menoufia University. Her current research interests include computer vision, data fusion, image processing, signal processing, medical imaging, modeling, and control.

ABEER D. ALGARNI received the B.Sc. degree (Hons.) in computer science from King Saud University, Riyadh, Saudi Arabia, in 2007, and the M.Sc. and Ph.D. degrees from the School of Engineering and Computer Sciences, Durham University, U.K., in 2010 and 2015, respectively. She has been working as an Assistant Professor at the College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, since 2008. Her current research interests include networking and communication systems, digital image processing, digital communications, and cyber security.

NAGLAA F. SOLIMAN received the B.Sc., M.Sc., and Ph.D. degrees from the Faculty of Engineering, Zagazig University, Egypt, in 1999, 2004, and 2011, respectively. She worked at the Faculty of Computer Science, PNU, Saudi Arabia. Her current research interests include digital image processing, information security, multimedia communications, medical image processing, optical signal processing, big data, and cloud computing.

RABIE A. RAMADAN (Member, IEEE) graduated from Alazhar University. He received the dual master's degrees from Cairo University and Southern Methodist University (SMU), Dallas, TX, USA, and the Ph.D. degree from the Computer Science and Engineering Department, SMU. He currently works as a Full Professor at the Computer Engineering Department, Cairo University, and is on leave from the College of Computer Science and Engineering, Hail University, Hail, Saudi Arabia. He has led many of the research projects in the fields of AI and smart technologies. He also works with many of the leading industrial partners around the world. He worked with Vodafone telecommunication operators, Intel, Samrtec, Microsoft, CERN, and CISCO Academy. He is the Founder of many AI research laboratories, including FABLAB Hail and Brain–Computer Interface (BCI) Laboratories. He also has a distinguished publication record in sensing technologies, including sensor networks and the Internet of Things (IoT). He is the Co-Founder of the Ambient Intelligent Center (AMIC), German University in Cairo (GUC). He also has a fruitful cooperation effort with the Fraunhofer Institute, Germany. He utilized AI techniques in security, personalized shopping, powered assistants, fraud prevention, administrative tasks automated to aid educators, creating smart content, voice assistants, personalized learning, and autonomous vehicles. He also secured funds for many research projects. He also maintained funding from different agencies for his projects and publications. He secured funds up to (U.S. $30,742,859) from different organizations, including NTRA, FP7, ITEA, Hail University, and MTI University.