applied sciences

MDPI

Article

Reverse Image Search Using Deep Unsupervised Generative Learning and Deep Convolutional Neural Network

Aqsa Kiran [1,2,3,4], Shahzad Ahmad Qureshi [2], Asifullah Khan [2,3,4], Sajid Mahmood [1,*], Muhammad Idrees [5,*], Aqsa Saeed [3], Muhammad Assam [6], Mohamad Reda A. Refaai [7] and Abdullah Mohamed [8]

[1] Department of Informatics and Systems (INFS), University of Management and Technology, Lahore 54000, Pakistan; aqsa.kiran@umt.edu.pk

[2] Department of Computer and Information Sciences (DCIS), Pakistan Institute of Engineering and Applied Sciences, Islamabad 45650, Pakistan; drsaqureshi@pieas.edu.pk (S.A.Q.); asif@pieas.edu.pk (A.K.)

[3] PIEAS Artificial Intelligence Centre, Pakistan Institute of Engineering and Applied Sciences (PAIC), Islamabad 45650, Pakistan; aqsa.qureshi@helsinki.fi

[4] Deep Learning Lab, Centre for Mathematical Sciences, Pakistan Institute of Engineering and Applied Sciences (PAIC), Islamabad 45650, Pakistan

[5] Department of Computer Science and Engineering, University of Engineering and Technology Lahore, Narowal Campus, Islamabad 54400, Pakistan

[6] Department of Software Engineering, University of Science and Technology Bannu, Bannu 28100, Pakistan; soft.researcher12@gmail.com

[7] Department of Mechanical Engineering, Prince Sattam bin Abdulaziz University College of Engineering, Alkharj 16273, Saudi Arabia; m.rifaee@psau.edu.sa

[8] Research Centre, Future University in Egypt, New Cairo 118355, Egypt; mohamed.a@fue.edu.eg

* Correspondence: sajid.mahmood@umt.edu.pk (S.M.); midrees10@uet.edu.pk (M.I.)

check for updates

Citation: Kiran, A.; Qureshi, S.A.; Khan, A.; Mahmood, S.; Idrees, M.; Saeed, A.; Assam, M.; Refaai, M.R.A.; Mohamed, A. Reverse Image Search Using Deep Unsupervised Generative Learning and Deep Convolutional Neural Network. *Appl. Sci.* **2022**, *12*, 4943. https://doi.org/10.3390/app12104943

Academic Editor: Vincent A. Cicirello

Received: 13 April 2022
Accepted: 4 May 2022
Published: 13 May 2022

Abstract: Reverse image search has been a vital and emerging research area of information retrieval. One of the primary research foci of information retrieval is to increase the space and computational efficiency by converting a large image database into an efficiently computed feature database. This paper proposes a novel deep learning-based methodology, which captures channel-wise, low-level details of each image. In the first phase, sparse auto-encoder (SAE), a deep generative model, is applied to RGB channels of each image for unsupervised representational learning. In the second phase, transfer learning is utilized by using VGG-16, a variant of deep convolutional neural network (CNN). The output of SAE combined with the original RGB channel is forwarded to VGG-16, thereby producing a more effective feature database by the ensemble/collaboration of two effective models. The proposed method provides an information rich feature space that is a reduced dimensionality representation of the image database. Experiments are performed on a hybrid dataset that is developed by combining three standard publicly available datasets. The proposed approach has a retrieval accuracy (precision) of 98.46%, without using the metadata of images, by using a cosine similarity measure between the query image and the image database. Additionally, to further validate the proposed methodology's effectiveness, image quality has been degraded by adding 5% noise (Speckle, Gaussian, and Salt pepper noise types) in the hybrid dataset. Retrieval accuracy has generally been found to be 97% for different variants of noise

Keywords: reverse images search; deep convolutional neural network; unsupervised representational learning; deep generative learning; sparse auto-encoder; ensemble learning; image retrieval

1. Introduction

The reverse image search is an emerging research area that overrides the conventional way of information retrieval, i.e., text-based search. The hosting of more than two billion images during 2004–2007 [1] is claimed by the image hosting online platform Flicker, and it is almost doubling every year, stressing the need for image search. Two types of image searches are done: one using the semantics of an image called text-based image retrieval, and the example-based visual search characterized by a lack of search terms. To label such a large image repository requires a lot of time, making the text-based image search a cumbersome task. Thus, covering the semantic gap by effectively removing the need for a user to guess that a result is based on relevant feedback of keywords or terms may or may not return a correct result. This visual search also allows users to discover content that is related to a specific sample image, the popularity of an image, and discover manipulated versions and derivative works [2]. The success of this representation is gaining attention [3] significantly in numerous fields such as image search engine [3], grouping and filtering of web images [4], in biomedical information management [5], and computer forensics and security [6]. Therefore, many researchers from different fields of science have focused their attention on image retrieval methods based on the visual content of images [7]. The reverse image search, also known as Content-Based Image Retrieval (CBIR), I, is an application of computer vision that searches relevant images from large databases. One of CBIR's main concerns is to lower the memory, which is required to store each image. A lot of research work has been carried out to develop more reliable and efficient search systems [8]. Some of the related work in the context of CBIR [9] as given below focuses mainly on the feature extraction techniques. Calentado et al. [10] presented a similarity evaluation using handcrafted simple feature extraction methods such as Hough transform, different position, and orientation of low-level image contents in HVC color space. Another simplified approach to some classification tasks that predicts the retrieved image category involves high overhead using wavelet transform techniques, Gabor filter-based wavelet transform, color correlogram, etc. [11]. Similarly, Kumar et al. [12] proposed a method for reverse image search based on grayscale images, which helped enhance the computational efficiency compared to the RGB images. Most of the research work used by the CBIR method utilizes a grayscale weighted system to reduce the characteristic vector dimensions. Grayscale is more suitable for color and texture image features' analysis compared to the color-weighted natural system. These methods use two common benchmark datasets, particularly Wang and Amsterdam Library of Texture Images (ALOT), to show the effectiveness of the approaches [13]. Previously, some limited work in the context of reverse image search has been done from the application point of view. Mistry et al. [2] have investigated CBIR procedures and their utilization in different application areas. In another work, Das et al. [14] have stressed color features using diverse components of images through four distinct strategies. Two of the four strategies were based on the analysis of color features, while the other two analyzed color and texture features. The color features alone have not been found reliable in case of image acquisition under poor lighting conditions. Malini et al. [15] proposed normal mean-based procedure with reduced features' size in combination with color averaging to accomplish higher retrieval effectiveness and execution rate. Tang et al. [16] introduce a new distance metric-based learning algorithm, namely weakly supervised deep learning, for image retrieval exploiting knowledge from community-contributed images associated with user-provided tags. Due to the success and powerfulness of deep learning (DL), several studies have been reported using deep learning approaches for image retrieval tasks [17]. This paper encapsulates such research works including some state-of-the-art research in the context of CBIR as well as methodologies comprising of ensembles of models for the sake of feature extraction and classification. Tefas et al. [18] produced compact feature vectors using convolutional vectors and convolutional layer activation regions. A mixed version of static and dynamic techniques was explored by Mohedano et al. [19] using a pipeline of CNN-features and the bag-of-words collection. Similarly, Yu et al. [20], proposed the exploitation of complementary strengths of CNN features in different layers. Another recent work done to reduce space consumption has been done by Ramzan et al. [21]. In Ramzan approach, the concept of bilinear CNN-based architecture in the CBIR domain is introduced, where a bilinear root pooling is proposed to project the features extracted from two parallel CNN models into a dimensionally reduced space. Besides all work done in the context of reverse image search or CBIR, there is still a need for a more accurate and reliable search system exploring variants of ML and DL techniques. Another recent work by Simran et al. [22] includes a straightforward but influential deep learning system focused on Convolutional Neural Networks (CNN) and comprised of feature extraction and classification for a firm image retrieval task.

However, most of the work in CBIR using DL is introduced in combination of two or more models such as collaboration of CNN variants with other models and shows considerable success compared to using standalone DL methods where the reliability and generalization ability become a question. For instance, in a work presented by Ouhda et al. [23], they design an approach with a convenient deep learning framework that ensembles Convolutional Neural Networks (CNN) with Support Vector Machine (SVM) to accomplish an efficient CBIR tasks. SVM is fed by the convolutional features coming from the CNN part initially. They ended their research by obtaining encouraging results among a pool of CBIR tasks using the image database. This typical trend of using CNN in combination with SVM was replaced by Pardede et al. [24], who put on the advantages of Deep CNN techniques and XGBoost classifier. The authors proposed a Deep CNN model to perform feature extraction and XGBoost as a classifier substituting the typical SoftMax and SVM classification. They observed the performance of Deep CNN for CBIR tasks generated using SoftMax, SVM, and XGBoost classifiers in terms of accuracy, precision, recall, and f1-score. Hence, these researchers claimed the enhanced performance by employing XGBoost classification experimenting upon Wang, GHIM-10k, and Fruit-360 image datasets. Similarly, Cui et al. [25] proposed a hybrid deep learning model based on deep Convolutional Auto-Encoder (CAE) complemented with CNN to solve this problem. Since the convolutional auto-encoder is well known for the provision of unsupervised feature extraction and data dimension reduction, so they utilized this concept for their objective of remote sensing data. Their model passed the CAE extracted features are presented as and to the following CNN, eventually classified to perform the retrieval results. This approach resulted in the final classification accuracy increase from 0.916 to 0.944, showing a considerable improvement upon simple solely CNN-based approaches. Lastly, another latest hybrid approach is introduced by Desai et al. [26] that comprises a framework using VGG-16 as a feature extractor and SVM to perform the final classification. They reported satisfying results encouraging more success towards CBIR by devising a combined methodology. Another stimulating fact is the use of generative learning in the exploration of supervised and unsupervised datasets. As highlighted by Dolikh et al. [27] and Abukmeil et al. [28], generative learning impacted high on the success and generalization ability of the computed feature space, especially the success of using unsupervised generative learning in the domain of computer vision. Another important study was done by Xie et al. [29], in which they familiarized a sparse framework by modifying an original image representational framework in order to develop a methodology that is loaded with high generative learning and can even be able to generate some realistic images of considerable quality.

Hence, by observing the impressive achievement in the image retrieval task by the deep-ensemble-based model, we break the stereotype of the standalone CNN model performing in this research work to develop a more effective and reliable CBIR methodology that basically addresses the major concern of improved generalization and classification ability of a reverse image search. The proposed research work introduces a new deep learning-based methodology in the context of CBIR, namely, Reverse Image Search using Deep Unsupervised Generative Learning and Deep Convolutional Neural Network (RIS-DUGL), which comprises a collaborative approach that ensembles two important models. It aims to improve the image search system's generalization and performance compared to the state-of-the-art reverse image search models developed. The proposed method boosts the faster convergence of reduced database results by defining a sparse representation. The proposed RIS-DUGL methodology consists of two steps. In the first step, a deep generative model is trained to get unsupervised representational learning and perform optimal parameter tuning. A Sparse Auto-Encoder (SAE) [30] customized with two layers that are tuned empirically is used. After getting a compact and efficient code, it is combined with ground truth channel wise. In the second step, the output of the first step is fed to a CNN variant, namely, (VGG-16) [31], to exploit transfer learning as a fixed feature extractor by using pertained ImageNet [32] database weights. Thus, it learns a unique image representation using a deep generative model. The retrieval task is carried out using cosine distance [33] as a similarity measure between the sample image feature vector and the feature database. Performance evaluation of the proposed RIS-DUGL technique is retrieval accuracy (precision), known as CBIR's most reliable measure. This research paper aims to introduce a solution in the context of reverse image search using dynamic and more powerful state-of-art techniques such as deep learning.

The proposed approach exploits unsupervised learning, which is underestimated in the context of CBIR. This paper is also useful as it covers both aspects of deep learning techniques, such as the one generative model by using unsupervised models, like auto-encoders, and the other powerful discriminative model using deep CNN, which has gained a lot of success in image processing and computer vision related tasks. The rest of the paper is organized as follows: Section 2 describes the proposed framework RIS-DUGL and its architecture. Section 3 discloses the two phases methodology proposed in RIS-DUGL along with the dataset knowledge and implementation details. It also describes how transfer learning is employed in the proposed method and highlights the variants of transfer learning. Section 4 discloses the implementation details and the experimental results. Section 5 concludes this research work, also mentioning the future direction of it.

2. Proposed Framework (RIS-DUG)

The framework mainly uses two deep learning models partitioned into two phases. The first phase of RIS-DUGL is initialized by exploiting unsupervised representational learning using a deep generative model without using any prior information about the dataset, as shown in Figure 1. We refer to phase-1 as the generative phase. The initialization in this phase assists in extracting useful information hidden in the original form of input. Moreover, to capture pixel-wise low-level details, each image is split into RGB channels before proceeding to phase-1. Each channel is sequentially forwarded to three SAEs, each of the SAEs in two layers deep to extract a useful and compact representations. The channel-wise output of SAEs generates three feature vectors that are concatenated with the respective original channels and then provided as input to phase-2, which we also refer to as the discriminative phase. Transfer learning is employed in this phase using a pertained, sixteen-layered deep VGG network. Finally, fine-tuning this pertained model is performed using custom-defined convolution and fully connected layers followed by their activation to compose the final feature database. The importance and intentions of transfer learning have been described in the feature extraction step in this section.

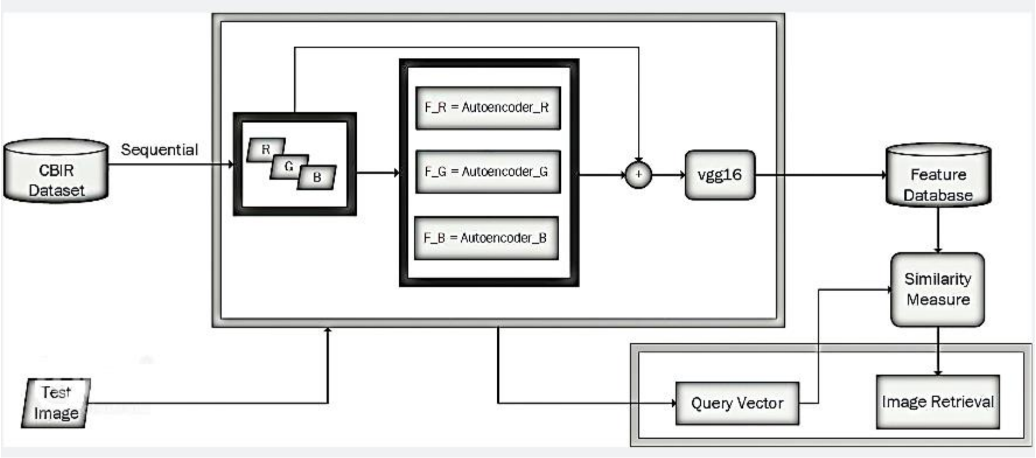


Figure 1. Architecture of the proposed RIS-DUGL retrieval model.

2.1. Dataset Description

A number of image databases exist these days for the development of reverse image systems, highlighting various information retrieval tasks. In the proposed RIS-DUGL technique, a hybrid image dataset by combining three publicly available image datasets is used.

2.1.1. Experiments on WANG Dataset

WANG-1000 test image dataset contains 10 classes having 100 instances of each class [34]. Some of the class instances have been shown in Figure 2. This dataset has widely been used for CBIR tasks specifically to perform comparisons.

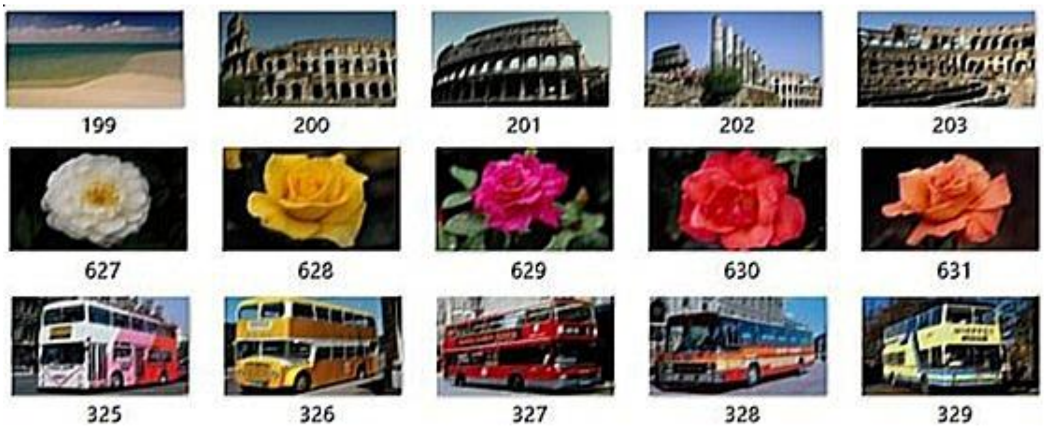


Figure 2. Sample images from the 10 classes of WANG-1000 image dataset.

2.1.2. Experiments on IAS-Lab RGB Face Dataset

This dataset is provided by the Intelligent Autonomous Systems Laboratory for computer vision related tasks [35]. It consists of sample images of 26 persons, each with 13 different poses with different light and expression conditions. Each sample image has been captured using a consumer sensor camera with (1920 × 1080) resolution. An essential aspect of this dataset is that various possibilities regarding the positions of subjects have been covered, as shown in Figure 3.

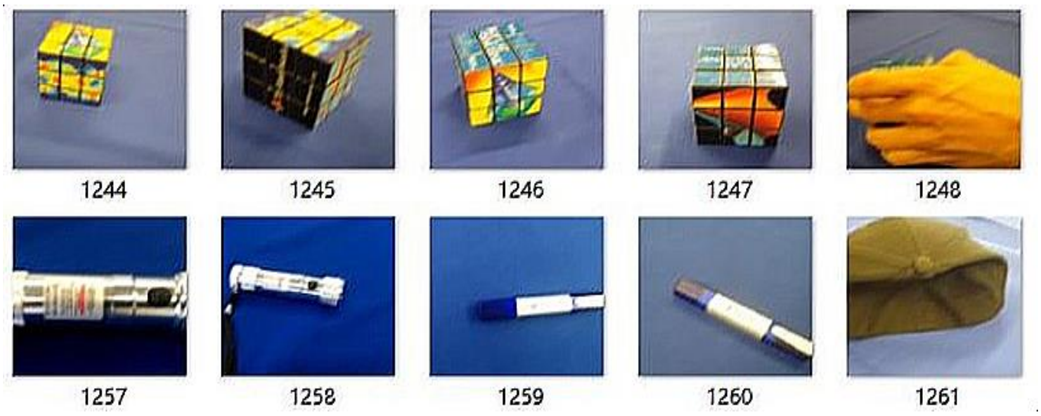


Figure 3. Preview of IAS-Lab RGB face dataset.

2.1.3. Experiments on Math Works Merchant Dataset

MATLAB is one of the most important mathematical computational software developed by Math Works Inc. [36]. It presents merchant dataset to perform basic learning tasks such as transfer learning [37]. This dataset is composed of twelve orientations and rotation invariant poses for six merchant objects such as screw driver, cube, play cards, torchlight, cap, etc., as shown in Figure 4.

Figure 4. Merchant dataset sample images.

2.2. Dataset Distribution

After selecting a proper dataset, its distribution and preparation is the most important step while developing a model. The training is performed on 80% of the hybrid dataset [38], while 20% of the dataset is test data and used to perform the experiment as shown in Figure 5. In this work, initially, the dataset has been split into training, validation, and test. The model is trained on the training set, while the validation set is used for validating the model parameters, keeping the testing data aside.

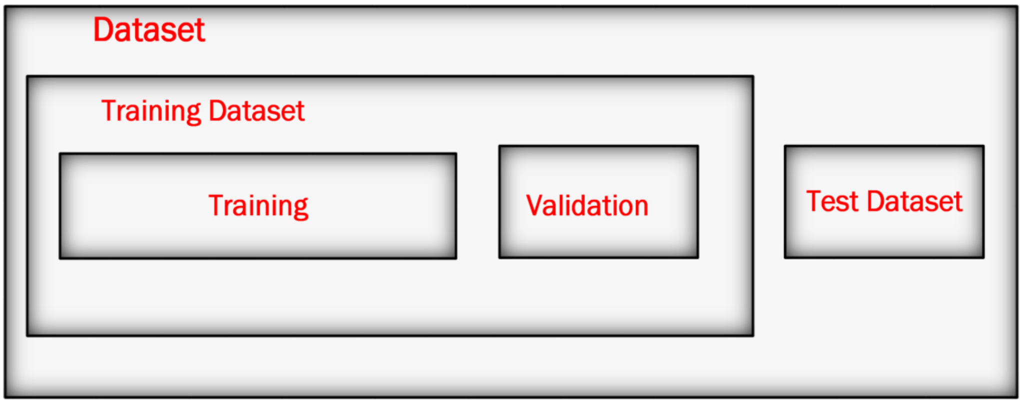


Figure 5. Block diagram for hybrid dataset distribution.

3. Ris-DUGL Methodology

Our approach framework consists of two phases, as shown in Figure 1. The two most important objectives are feature extraction and generation of an efficiently computed compact database. Image retrieval task is carried out for testing the model. The various parts of the model are given below:

3.1. Unsupervised Representational Learning Using Deep Generative Model

Supervised learning has the limitation of being dependent on the prior knowledge about the dataset, such as labels and annotations (in the case of images, this type of meta-data is the prior knowledge) to perform the classification task. On the other hand, the model of unsupervised learning is trained without any meta-data about the dataset. It is referred to when performing feature extraction, dataset generation, or input reconstruction using fewer dimensions to represent the input. It also performs clustering tasks based on similar representations. In image retrieval, we want to learn the underlying image representation, which is referred to as representational learning [39]. Thus, unsupervised representational learning is the art of learning the underlying structure and distribution of data without using labels or other information. This is a more challenging and more useful task, as this technique builds such a powerful model that is more effective and reliable by learning from self-mistakes. There exist various deep learning-based generative models that perform unsupervised representational learning [40]. One such powerful model is Auto-encoder (AE). It is a neural network that copies the input to its output using fewer dimensions in an unsupervised manner [41]. An AE architecture consists of the encoder that helps in learning and converting input into a hidden representation and a decoder, which is used for reconstructing the original input from the encoded representation as described by Equations (1) and (2):

$$ien = fen(iorig) \tag{1}$$

$$ide = fde(ien) \tag{2}$$

where $iorig$ is the original input taken by encoding function fen and returns encoded representation ien. Similarly, fde is the function the reconstructs the original input ien as ide.

Copying an input to its output is not an effective and desired task, as it only memorizes what it has seen. This memorization may lead to over fitting, resulting in poor generalization. A variant of AE exists with more generalization power, namely, sparse auto-encoder (SAE) [42]. It copies the input data, attempts to learn the real underlying data distribution, and performs accurately on the test data. SAE focuses on learning the data distribution by restricting the number of connections in the hidden layer [39,40]. An AE architecture consists of an input layer, an output layer, and a single hidden layer responsible for the encoding function of the input data, as shown in Figure 6. An under complete AE has a smaller number of neurons in the hidden layer than the input layer. The number of encoding neurons in each hidden layer is equal, but they fire their decision using the sparsity elements, making the architecture of SAE unique, as shown in Figure 7. SAE has the choice to activate neurons of the network within each hidden layer selectively, thereby introducing sparsity in the connections between each layer [43]. Compared to the simple AE version, SAE constraints the network's capacity to memorize the input data without limiting its capability to learn the underlying representation of the input data. In SAE, loss reduction is based on using the mean-squared-error (MSE) along with the sparsity conditions. To impose these conditions, the loss function of SAE is given by Equation (3):

$$L = \frac{1}{N}\sum_{x=1}^{M}\sum_{y=1}^{N}\left(X_{xy} - \overline{X_{xy}}\right)^2 + \alpha * \Omega_{wr} + \beta * \Omega_{sr}, \tag{3}$$

where $\frac{1}{N}\sum_{x=1}^{M}\sum_{y=1}^{N}\left(X_{xy} - \overline{X_{xy}}\right)^2$ is the MS. The other terms Ω_{wr} *and* Ω_{sr} are L2 weight regularization and sparsity regularization, respectively, along with their co-efficients α and β to control their impact. L2 weight regularization adjusts the influence of the weights of the network, recommended to be a smaller value and defined in terms of weights as given by Equation (4):

$$\Omega_{wr} = \frac{1}{2}\sum_{z'}^{H}\sum_{x'}^{M}\sum_{y}^{N}\left(W_{xy}^{z'}\right)^2 \tag{4}$$

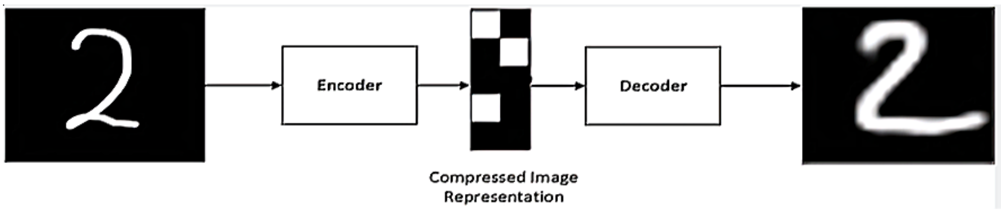


Figure 6. Illustration of auto-encoder functionality.

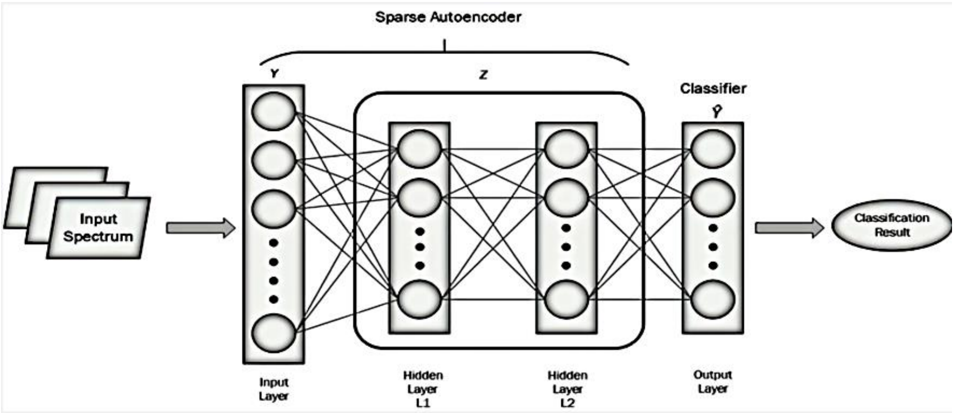


Figure 7. Architecture of deep-sparse auto-encoder (SAE).

Here, H, M, and N are the number of hidden layers, the number of examples, and the number of actually used variables in the data, respectively. This L2 weight regularization is the major source for good generalization. Another constraint, Sparsity Regularization, helps to panelize the sparsity of the output connections from the hidden layer as given by Equation (5):

$$\Omega sr = \sum_{x=1}^{s} KLD'\left(\frac{q}{qx}\right) = \sum_{x=1}^{s} KLD'\left(\frac{q}{qx}\right) + (1-q)\log\left(\frac{1-q}{1-qx}\right) \quad (5)$$

where q is the desired value for neuron x, and qx is the average value for neuron x. The difference between the actual and the desired values, i.e., between q and qx of the neuron, x increases the value of Ωsr. Another important parameter is the scarcity proportion, which is adjusted within the scarcity regularization. It controls the scarcity of the output from the hidden layer. To specialize each neuron in a layer, by choosing a value from the low range, only giving a high output for a small number of training examples. For example, if the scarcity proportion is set to 0.1, this is equivalent to saying that each neuron in the hidden layer should have an average output of 0.1 over the training examples. This value must be between 0 and 1. The ideal value varies depending on the nature of the problem. A specific range of values is explored in the proposed RIS-DUGL technique with scarcity regularization, i.e., α = 6 and scarcity proportional = 0.002. However, the possible loss reduction in SAE was achieved by the optimal parametric tuning. After performing preprocessing like calling and resizing on the image (256 × 384 × 3), each image is split into its RGB channel and every channel fed to a two layers deep SAE [40] in a sequential manner, as shown in Figure 7. FR, FG and FB are three feature vectors using SAE, as shown in Figure 1. SAE has fewer connections, which can be maintained by adjusting a scarcity proportion. A major variation introduced in this variant of the simple auto-encoder is that it regularizes the loss by adding some penalties to learn the best representation. The second major concern to use SAE is that it learns sparse representation and helps us learn highly discriminative compact features. Each feature vector of the compact code is concatenated with the corresponding original RGB channel. The resulting three-dimensional matrix is prepared as input to the second phase.

3.2. Feature Extraction Using Deep CNN

Convolutional Neural Network (CNN) is an Artificial Neural Network (ANN) that uses convolutional operations and has fewer connections. CNNs consists of convolution layer, pooling (sub-sampling) layers, and fully connected layer followed by the output layer, as shown in Figure 8. Each convolutional layer gets random parameters initialized with the filter or neuron values. It may get starting values also from the pre-trained model [44]. There exist various architectural variations in deep convolutional neural networks, enhancing their capabilities. Generally, a max-pooling layer is a part of CNN architecture, as it summarizes the outputs of neighboring groups of neurons in the same kernel map.

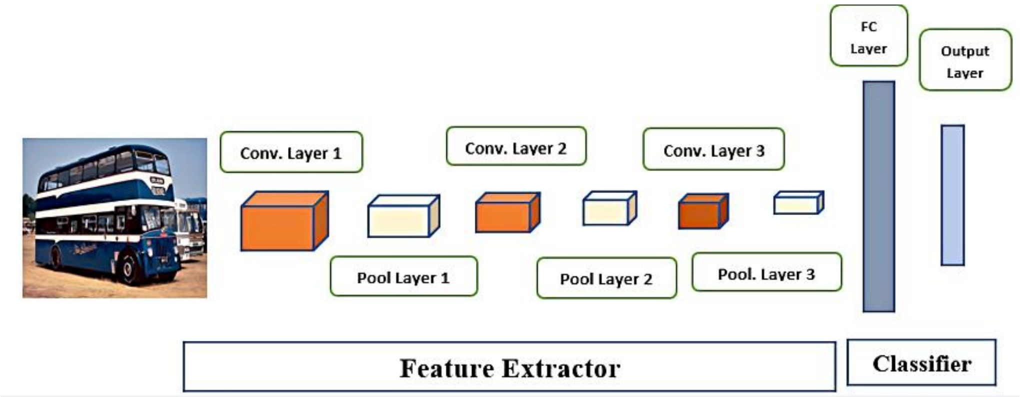


Figure 8. Overview of CNN architecture.

In the proposed work, transfer learning is employed to extract the final features using a recent variant of CNN architecture, namely, VGG-16 [31]. Figure 9 shows the architecture of VGG-16, which consists of sixteen layers, including convolutional pooling layers, fully connected layers, followed by the output layer. In the proposed technique, a pre-trained model on ImageNet [45] is used, which consists of one thousand classes of images with one million images per class. It saves us from the difficulty to train VGG-16 from scratch. The fully connected layer is discarded to extract all features directly from the convolutional feature maps' activation before the last layer. This step feeds a three-dimensional matrix (224 × 224 × 3) and computes a 4096-D vector on each image. An important step is to apply an activation function. We used one of the most powerful differential functions, namely, the rectified linear unit (ReLU), which thresholds the values at zero. The need to do this step because each layer's activation is also a threshold during the training of the network on ImageNet.

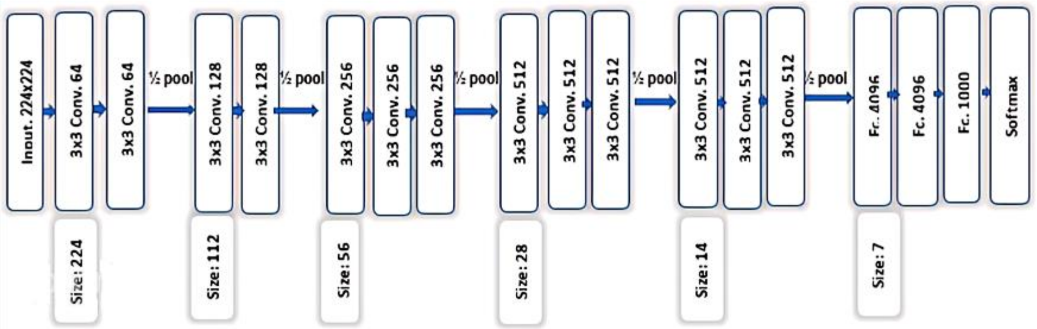


Figure 9. VGG-16 architecture.

Another successful variant of CNN is Alex-Net [44], which is an eight-layer deep architecture, as shown in Figure 10. It is most widely used to perform feature extraction in computer vision-related applications. In this work, the exploitation and identification of the proposed methodology's learning ability and competence are carried out. The mechanism behind the feature extraction using Alex-Net is the same as mentioned for VGG-16. By using multiple deep learning models, this research highlights the generalization and flexibility of the proposed reverse image search system, i.e., we do not constrain the model. An important concept in the deep learning technique is to adopt already trained models for improving performance-related tasks, usually referred to as transferring knowledge of similar models, known as Transfer Learning (TL).

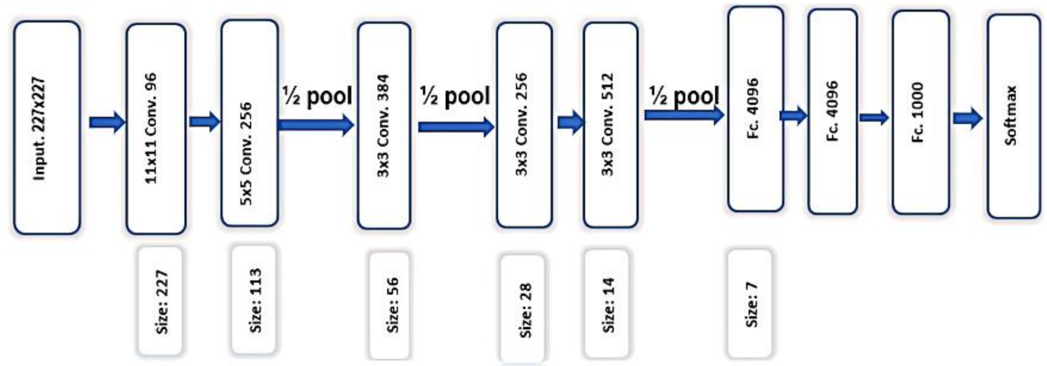


Figure 10. Alex-Net architecture.

3.3. Transfer Learning

It produces noteworthy results when the source and target tasks are interconnected; otherwise, the target task's performance may not be promising. This concept is useful for training deep learning architectures because they consume more computational time during their training [44,46]. Figure 11 shows the concept of TL that saves us from the difficulty of training a model from scratch by using an already designed method that has been trained on a similar data distribution. Thus, TL greatly reduces the cost of designing similar models and pays the cost to get new training data each time by using the knowledge learned in one model on a similar dataset for other related tasks. It is specifically desirable to deal with a smaller number of training instances. Generally, the researchers use the pre-trained original deep neural network on their customized dataset. After that, fine-tuning of the learned features is done for another dataset with a new target network, as shown in Figure 12. For example, a model trained for different dog classes can also be used to classify cats by performing fine- tuning. TL is employed in the second phase in the proposed work by using a pre-trained deep discriminative model, which is trained on similar and larger image domains. It results in an improvement in generalization. There exist different scenarios, which describe the use of Transfer Learning, categorized as follows in the sub-section:

3.3.1. Use of Transfer Learning in the Proposed RIS-DUGL Technique

A pre-trained deep learning model is used in this type, which is trained on a larger database, usually on ImageNet, and uses the activations of fully connected layers as fixed features. In other words, the classification is discarded, and features of the new dataset are extracted, as shown in Figure 12. One of the major benefits of this type is that it performs well even if there is an insufficient dataset available in the target domain. Moreover, it reduces the risk of over fitting as well [37,44]. Thus, without training our model from scratch, useful features can be easily extracted using TL concepts.

3.3.2. Transfer Learning with Fine-Tuning the Pre-Trained Model

To use Transfer Learning to perform the classification task, there is no need to discard the last fully connected layer. The output layer in the original model that consists of the Soft-Max classifier function to fire the decision is there, but the classifier is replaced, and the model is retrained using the target dataset. This is known as fine-tuning [46], illustrated in both Figures 11 and 12. However, it depends on whether we want to retain all layers of the model to perform fine-tuning, or we can freeze the initial layers as they learn almost the same parameters. Usually, the difference occurs at the later ones.

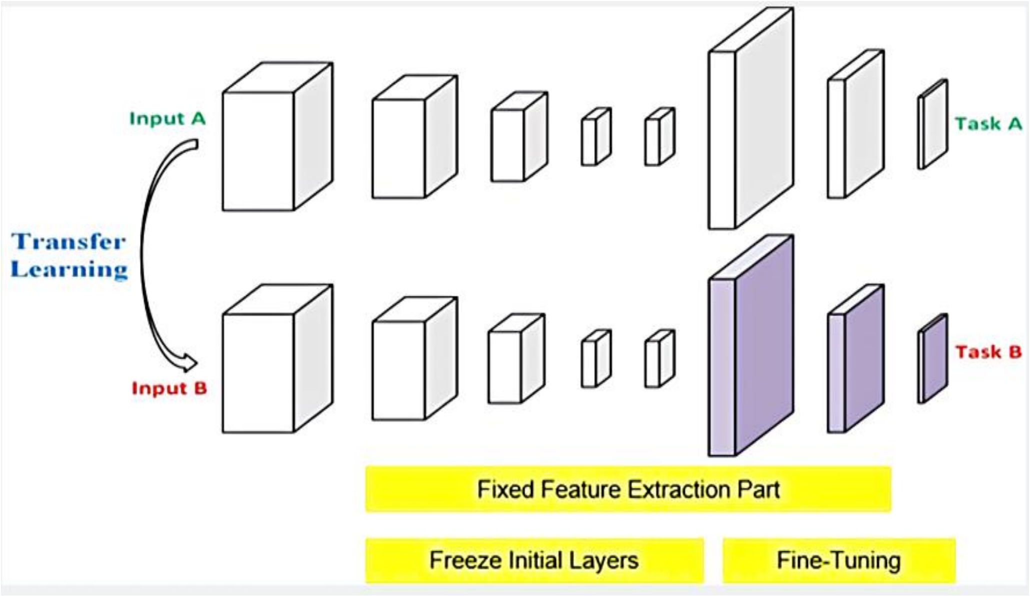


Figure 11. Overview of transfer learning.

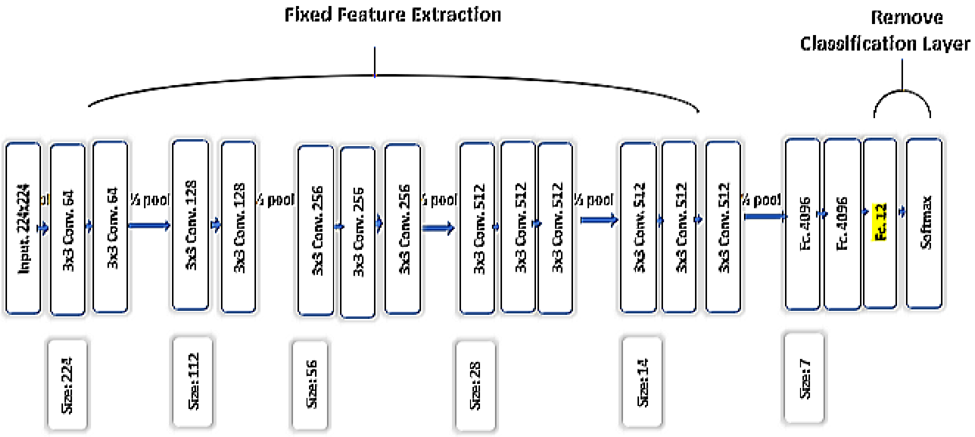


Figure 12. Use of transfer learning as a fixed feature extraction.

3.4. Retrieval Task and Performance Evaluation

The task of the proposed RIS-DUGL methodology is to find the best match to the query image. It is passed from both phases of the proposed RIS-DUGL technique and converted into a feature vector whose similarity is measured with each of the feature vectors in the feature database. To find the relevant match between the query vector and the feature vectors of the training database, various distance-based similarity matrices exist, such as Euclidean, Cosine, Manhattan, etc. A brief description is given below:

3.4.1. Cosine Similarity

This measure (the maximum value is unity) decides the best match to the sample image to perform the retrieval task. In the proposed work, both Euclidean and Cosine standard distance measures [47] have been used as an evaluation measure. Cosine similarity is considered generally the most reliable measure. The final evaluation of similar images in the proposed methodology is done using the cosine measure. Mathematically, a similarity score between the query vector q and the feature database di, i.e., $\sum_{i=1}^{\infty} = 1(di)$, the cosine distance measure is defined by Equation(6) as given by:

$$d_{cos}(q, d) = 1 - cos\theta = 1 - \frac{q \cdot d}{|q| \cdot |d|} \tag{6}$$

3.4.2. Retrieval Accuracy: Precision

Since the consequence of an information retrieval system specifically in RIS, the instances are pictures, and the task is to return a set of the most relevant pictures given a sample search image i.e., to assign each image to one of the two categories, “relevant” and “not relevant”. In this case, the “relevant” images are simply those having similar content to that of the desired object. Precision helps measure the number of relevant images retrieved against a RIS providing the total number of images retrieved by that search. The ideal precision score is 1.0, which means that every result retrieved against the sample image is relevant. The precision of the system characterizes the retrieval accuracy to retrieve how much a similar match to the sample image, i.e., the content of the data will be available upon request and can be accessed in collaboration with the corresponding author. A retrieved image belongs to the same domain in which the test image lies. Precision is known to be the best measure for the retrieval evaluation performance of CBIR [47]. It is the ratio of retrieved examples that are relevant in the whole database, known as the true positive (TP) and the sum of true predicted instances (TP) and falsely predicted instances (FP) retrieved as a result of RIS, as given by in Equation (7):

$$Precision\ (P) = \frac{TP}{TP + FP} \tag{7}$$

The primary objective of any reverse image search system is to present the most related images satisfying the user query image. The reason is that to find the exact match without any irrelevant extracted results is not possible generally.

3.5. Implementation Details

Final experiments of RIS-DUGL methodology are carried out on the desktop machine with a 3.4GHz processor clock and 4GB RAM. This desktop machine’s operating system is windows-10 Professional Edition, and MATLAB R2018a [39] has been used as the programming tool. A specific range of values is explored in the proposed RIS-DUGL technique manually. However, the possible loss reduction in SAE was achieved by the optimal parametric tuning. A specific range of values for the hyper parameters of deep SAE is explored in the proposed RIS-DUGL technique. The optimal values used are shown in Table 1. The outcome of SAE is channel wise concatenated with the original channel, and the final three-dimensional value is passed to VGG-16.

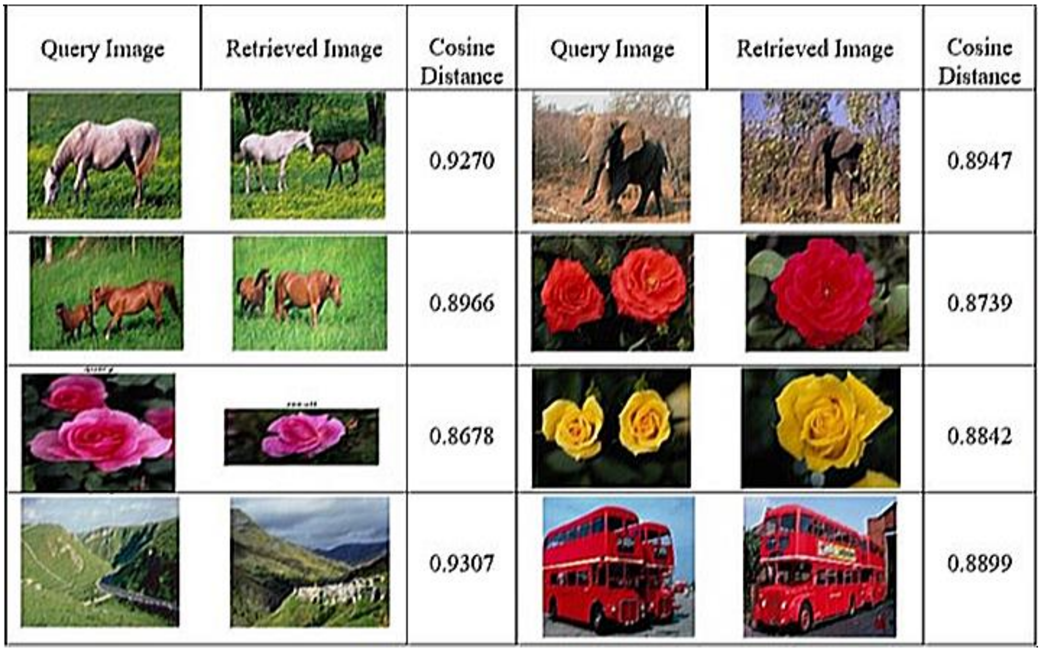
Table 1. Parameter values of Sparse Auto encoder used in the proposed RIS-DUGL.

No. of Layers	**1**	**2**
No. of Neurons Per Layer	100	80
L2 Weight Regularization	0.001	0.001
Sparsity Regularization	6	5
Sparsity Proportion	0.1	0.1
Scale	True	True
Epochs	50	20

4. Experimental Results

During the experiment, test data is used to evaluate the performance of the proposed technique. RIS-DUGL method does not use any prior knowledge such as labels (except those used for the fine-tuning) and other metadata of images. Results related to WANG, IAS- RGB, and Merchant datasets are shown in Figures 13–15, respectively, along with their cosine similarity scores. With various conditional query images such as left, down, and right side poses of faces. It also covers the image’s sufficient depth to recognize the subject lying far from the sample image for the retrieval of its more visible version. In

Figure 13, some of the results collected on the WANG dataset and the proposed framework have recognized the desired subject on the sample image having multiple subjects. It also covers color manipulation and recognizes the desired subject with background conditions. In Figure 14, it can be seen that some of the query images have equal participation in the desired object and the background information. Here, the proposed RIS-DUGL technique gives more value to the content of the desired subject and finds out its more visible version, which is a prominent feature for the sake of person identification. Figure 15 shows results using the Merchant dataset where the retrieved scale and rotation invariant best match. By using precision as a performance measure, the proposed RIS-DUGL technique has achieved 98.46% accuracy.



Query Image	Retrieved Image	Cosine Distance	Query Image	Retrieved Image	Cosine Distance
		0.9270			0.8947
		0.8966			0.8739
		0.8678			0.8842
		0.9307			0.8899

Figure 13. Results on WANG 1000 images.

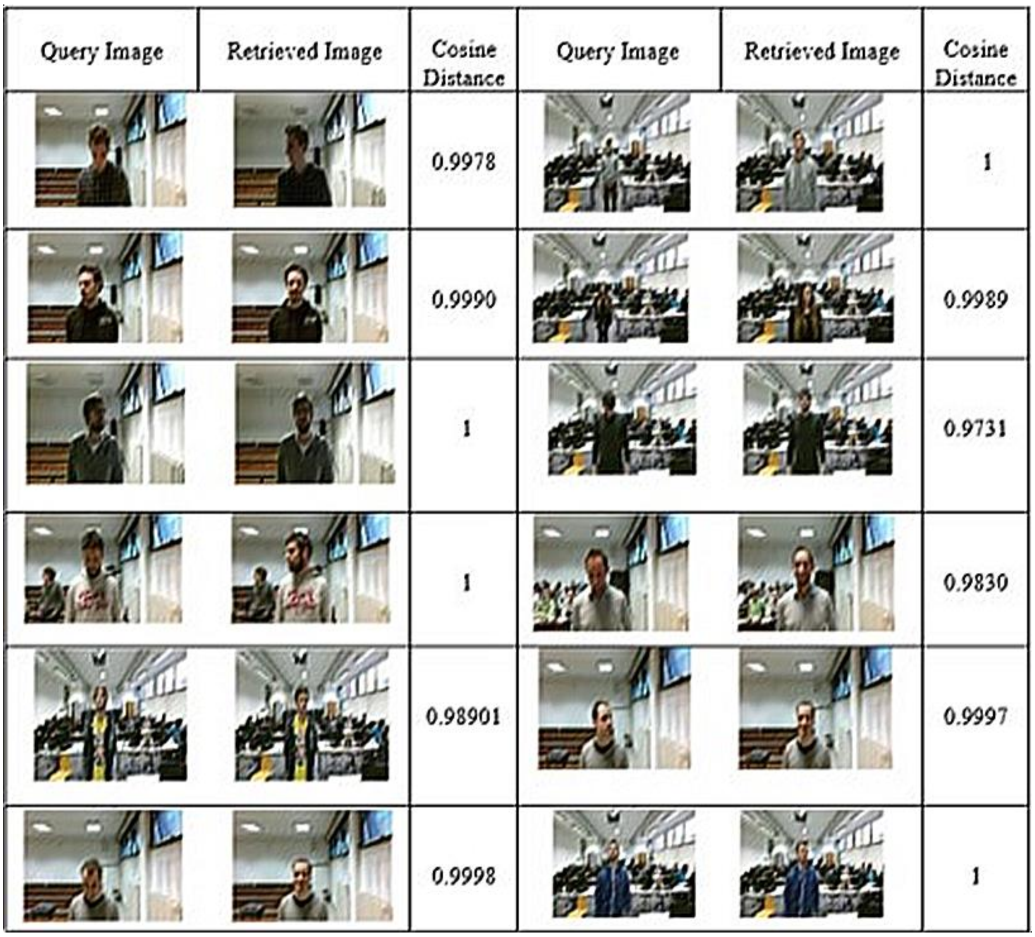
4.1. Comparison with Conventional Methods

Since there exist powerful static image processing techniques like Hough Transform, Gabor Filter, Discrete Wavelet Transform, etc. [10,11], that have been utilized considerably in classical research work in the context of reverse image search, we started out the study on some well-known classical research works and presented them for the sake of comparison with our proposed (RIS-DUGL) work's performance. We have implemented two of the conventional approaches, "Similarity Evaluation in Image Retrieval using Simple Features" [10] and "Content-Based Image Retrieval using SVM, NN, and KNN Classification" [14] in the context of CBIR. Table 2 summarizes and summarized comparison. These approaches use static digital image processing techniques. These techniques use shallow and hand-engineered fixed parameters and use simple features for content information from raw images and then use some similarity evaluation to do a reverse image search. Hence, the proposed RIS-DUGL method outperforms the retrieval accuracy compared to these two conventional techniques, as summarized in Table 2.

4.2. Results and Discussion on Noise-Induced Hybrid Dataset Using RIS-DUGL

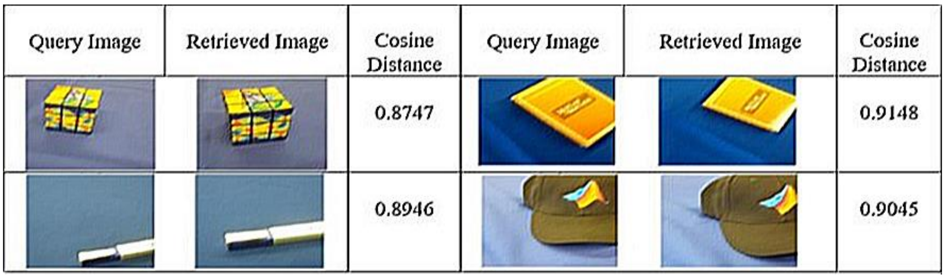
A comprehensive noise study has also been carried out to check the proposed reverse image search system's robustness. Three types of noise, namely, Salt & Pepper, Speckle, and Gaussian-noise, have been used to impregnate the original dataset. Salt & Pepper noise is an artifact that has a dark intensity value in bright areas and bright intensity values in dark regions. This type of noise can be triggered by analog-to-digital conversion. Gaussian noise is nothing but a random valued noise that follows a uniform or other distribution. It can have any random value of random variables [48].

A sample image with noisy ones has been shown in Figure 16. In this work, noise is added with a 0.05 ratio to corrupt the original dataset. Retrieval accuracy for noisy data implementation has been found to be 97%, 96%, and 97% using Gaussian, Salt & Pepper, and Speckle noise, respectively. This clearly indicates the effectiveness and high capability of the proposed method in the case of using dynamic feature extraction. Hence, we can announce that dynamic feature extraction (using Deep Learning) proves to have more learning ability and focuses more on generalization. Figure 17 summarizes the experimental results when the RIS-DUGL dataset is trained and tested on noisy images, starting with the speckle noise in the first query and retrieved images followed by Gaussian and Salt & Pepper noise. For noisy image retrieval, starting from the first sample image of a human with a face-up position, which includes Speckle noise, the proposed method retrieves the exact human with its standing position. This clearly indicates the importance and high capability of this work for criminal detection. In the case of the horse sample image, the image retrieved has the best match of the brown horse having a quite different pose and a distant object in the sample image. Clearly, when tested with the flower sample image containing Salt &Pepper noise, the proposed RIS-DUGL technique returned its image to the same flower class with many instances, and the manipulated color is also identified. In the case of a rotationally transformed cap, a sample image containing Gaussian noise, the similar orientation-wise transformed image is retrieved as the best image. The optimal training performance of RIS-DUGL is attained by using a maximum of 25 iterations for two layers of deep SAE, as can be seen from Figure 18, at the 25th iteration, the learning performance gets saturated.



Query Image	Retrieved Image	Cosine Distance	Query Image	Retrieved Image	Cosine Distance
		0.9978			1
		0.9990			0.9989
		1			0.9731
		1			0.9830
		0.98901			0.9997
		0.9998			1

Figure 14. Results on IAS-RGB Lab dataset.



Query Image	Retrieved Image	Cosine Distance	Query Image	Retrieved Image	Cosine Distance
		0.8747			0.9148
		0.8946			0.9045

Figure 15. Results on Merchant dataset.

Table 2. Comparison of time and accuracy of the proposed model with conventional methods.

Methodologies	Retrieval Accuracy (Precision)	Average Feature Extraction Time per Image (s)	Average Retrieval Time per Image (s)
CBIR with SVM based Classification [14]	39.24%	5.23	6.79
Similarity Evaluation using Simple Features [10]	74.42%	0.31	0.32
Proposed RIS-DUGL (AlexNet)	94.56%	0.23	0.28
Proposed RIS-DUGL (VGG-16)	98.46%	0.50	0.23

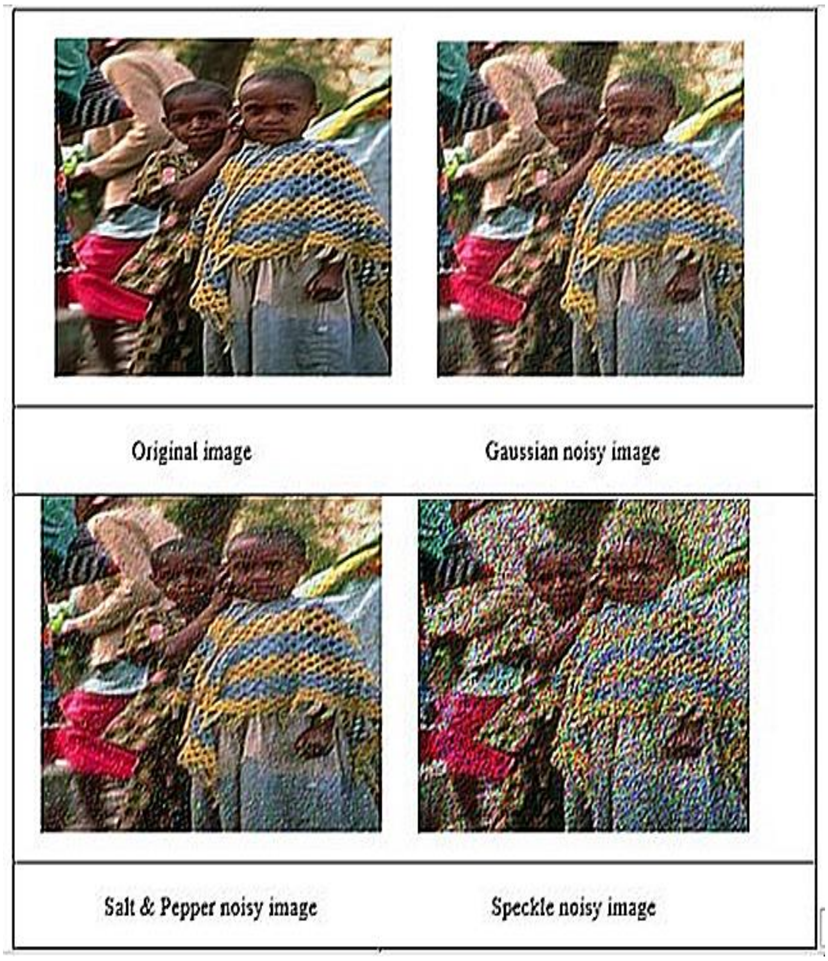


Figure 16. Sample images corrupted with three types of noise.

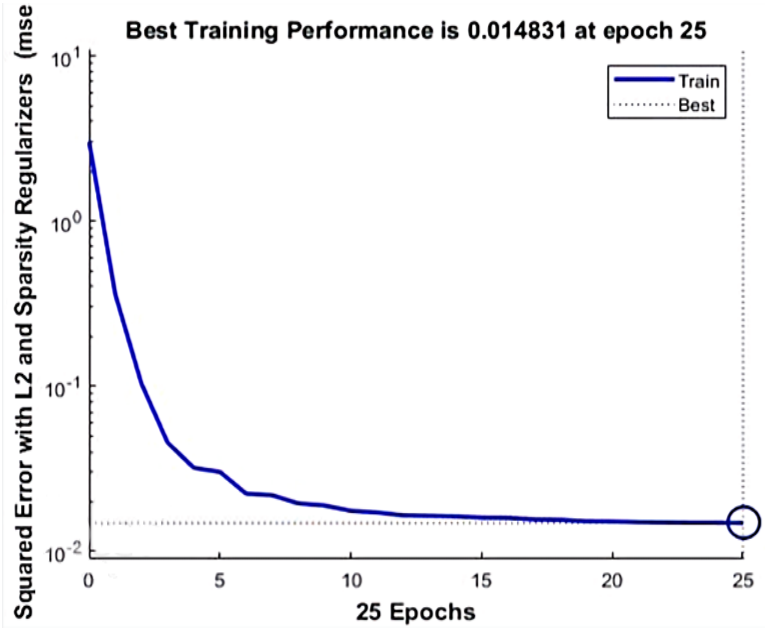
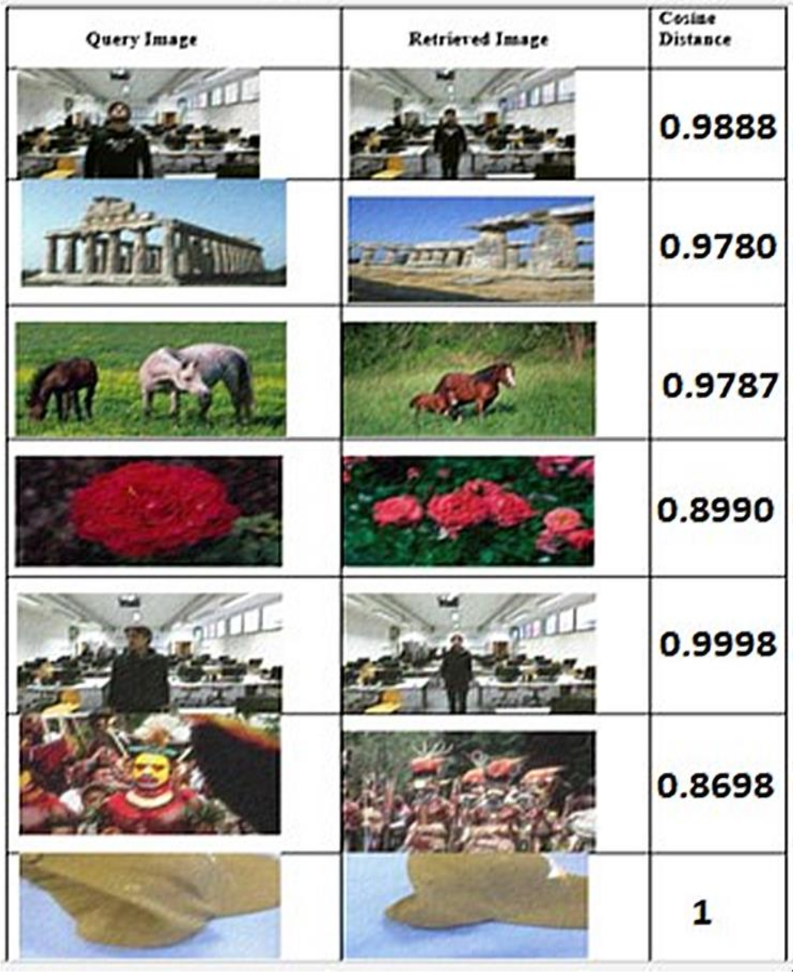


Figure 17. Results of RIS-DUGL using noisy hybrid dataset.



Query Image	Retrieved Image	Cosine Distance
		0.9888
		0.9780
		0.9787
		0.8990
		0.9998
		0.8698
		1

Figure 18. Training process of Sparse Auto-encoder (SAE) in the proposed.

5. Conclusions and Future Recommendation

A novel two phase reverse image search methodology, namely RIS-DUGL, based on deep learning techniques and transfer learning, is proposed. RIS-DUGL uses two deep neural network models sequentially. The first phase works in an unsupervised manner by using one of the deep generative models called Sparse AE and performs representational learning. In the second phase, the final feature extraction is done using a deep convolutional neural network called VGG-16. The proposed method exploits RGB channel-wise information, which provides rich representational learning and enhances the generalization ability by overcoming various content-based image retrieval (CBIR)/digital image challenges. RIS-DUGL practically employs the effective transfer learning approach by fine-tuning the already trained similar model to enhance the efficiency and reduce the trouble of training a model from scratch. This research activity is evaluated by a hybrid dataset comprising of 12 diverse categories collected from three publicly available datasets and using the most stable cosine distance as a similarity measure. Experiments reported 98.46% accuracy. To check the model's effectiveness, 5% noise (Gaussian, Salt &Pepper, and Speckle) is added in the hybrid dataset, which reported 97%, 96%, and 97% retrieval accuracy, respectively. The future direction is to adapt the trending Deep Learning Models (such as Generative Adversarial Networks (GANs)) to exploit further achievements in RIS-DUGL. Furthermore, a large-scale dataset, including many examples, may be used with training instances.

Author Contributions: Conceptualization, A.K. (Aqsa Kiran); data curation, S.A.Q.; formal analysis, A.K. (Asifullah Khan); funding acquisition, M.A. (Muhammad Assam); investigation, S.M.; methodology, M.I.; project administration, A.S.; resources, A.M.; supervision, M.R.A.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data will be available upon request and can be accessed in collaboration with the corresponding author.

Acknowledgments: The authors express their gratitude towards PIEAS Artificial Intelligence Centre (PAIC), DCIS: PIEAS, Department of INFS, UMT for providing infrastructure and a conducive environment for this research.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. Rafiee, G.; Dlay, S.S.; Woo, W.L. A Review of Content-Based Image Retrieval. In Proceedings of the 2017 International Symposium on Communication Systems, Networks & Digital Signal Processing (CSNDSP2010), Newcastle upon Tyne, UK, 21–23 July 2010; pp. 775–779.
2. Misty, Y.; Ingle, D. Survey on Content Based Image Retrieval Systems. *Int. J. Innov. Res. Comput. Commun. Eng.* **2013**, *1*, 1828.
3. Júnior, d.S.; Augusto, J.; Marçal, R.E.; Batista, M.A. Image Retrieval: Importance and Applications. In Proceedings of the Workshop de Visao Computacional-WVC, Uberlândia, MG, Brazil, 6–8 October 2014.
4. Wu, O.; Zuo, H.; Hu, W.; Zhu, M.; Li, S. Recognizing and Filtering Web Images based on People's Existence. In Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Sydney, NSW, Australia, 9–12 December 2008; Volume 1, pp. 648–654.
5. Kanumuri, T.; Dewal, M.; Anand, R. Progressive medical image coding using binary wavelet transforms. *Signal Image Video Processing* **2014**, *8*, 883. [CrossRef]
6. Brown, R.; Pham, B.; Vel, O.D. Design of a Digital Forensics Image Mining System. In *Proceedings of the International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 395–404.
7. Ranjan, R.; Gupta, S.; Venkatesh, K.S. Image retrieval using dictionary similarity measure. *SIViP* **2019**, *13*, 313–320. [CrossRef]
8. Alsmadi, M.K. Content-Based Image Retrieval Using Color, Shape and Texture Descriptors and Features. *Arab. J. Sci. Eng.* **2020**, *45*, 3317–3330. [CrossRef]
9. Alturki, R.; AlGhamdi, M.J.; Gay, V.; Awan, N.; Kundi, M.; Alshehri, M. Analysis of an eHealth app: Privacy, security and usability. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 209–214. [CrossRef]

10. di Sciascio, E.; Celentano, A. *Storage and Retrieval for Image and Video Databases*; International Society for Optics and Photonics: Bellingham, WA, USA, 1997; Volume 3022, pp. 467–477.
11. Das, S.; Garg, S.; Sahoo, G. Comparison of content-based image retrieval systems using wavelet and curvelet transform. *Int. J. Multimed. Its Appl.* **2012**, *4*, 137. [CrossRef]
12. Kumar, K.; Li, J.P.; Shaikh, R.A. Content based image retrieval using gray scale weighted average method. *Int. J. Adv. Comput. Sci. Appl.* **2016**, *7*, 1–6. [CrossRef]
13. Oğul, H. ALoT: A time-series similarity measure based on alignment of textures. In *International Conference on Intelligent Data Engineering and Automated Learning*; Springer: Cham, Switzerland, 2018; pp. 576–585.
14. Singh, S.; Rajput, E.R. Content based image retrieval using SVM, NN and KNN classification. *Int. J. Adv. Res. Comput. Commun. Eng.* **2015**, *4*, 549–552.
15. Malini, R.; Vasanthanayaki, C. An Enhanced Content Based Image Retrieval System using Color Features. *Int. J. Eng. Comput. Sci.* **2013**, *2*, 3465–3471.
16. Li, Z.; Tang, J. Weakly supervised deep metric learning for community-contributed image retrieval. *IEEE Trans. Multimed.* **2015**, *17*, 1989–1999. [CrossRef]
17. Asam, M.; Hussain, S.J.; Mohatram, M.; Khan, S.H.; Jamal, T.; Zafar, A.; Khan, A.; Ali, M.U.; Zahoora, U. Detection of exceptional malware variants using deep boosted feature spaces and machine learning. *Appl. Sci.* **2021**, *11*, 10464. [CrossRef]
18. Tzelepi, M.; Tefas, A. Deep convolutional learning for content based image retrieval. *Neurocomputing* **2018**, *275*, 2467–2478. [CrossRef]
19. Gomez Duran, P.; Mohedano, E.; McGuinness, K.; Giró-i-Nieto, X.; O'Connor, N.E. Demonstration of an open source framework for qualitative evaluation of CBIR systems. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Korea, 22–26 October 2018; pp. 1256–1257.
20. Yu, W.; Yang, K.; Yao, H.; Sun, X.; Xu, P. Exploiting the complementary strengths of multi-layer CNN features for image retrieval. *Neurocomputing* **2017**, *237*, 235–241. [CrossRef]
21. Alzu'bi, A.; Amira, A.; Ramzan, N. Content-based image retrieval with compact deep convolutional features. *Neurocomputing* **2017**, *249*, 95–105. [CrossRef]
22. Simran, A.; Kumar, P.S.; Bachu, S. Content Based Image Retrieval Using Deep Learning Convolutional Neural Network. In *IOP Conference Series: Materials Science and Engineering*; IOP Publishing: Bristol, UK, 2021; Volume 1084, p. 012026.
23. Mohamed, O.; Khalid, E.A.; Mohammed, O.; Brahim, A. Content-based image retrieval using convolutional neural networks. In *First International Conference on Real Time Intelligent Systems*; Springer: Cham, Switzerland, 2019; pp. 463–476.
24. Pardede, J.; Sitohang, B.; Akbar, S.; Khodra, M.L. Improving the Performance of CBIR Using XGBoost Classifier with Deep CNN-Based Feature Extraction. In Proceedings of the 2019 International Conference on Data and Software Engineering (ICoDSE), Pontianak, Indonesia, 13–14 November 2019; pp. 1–6.
25. Cui, W.; Zhou, Q. Application of a hybrid model based on a convolutional auto-encoder and convolutional neural network in object-oriented remote sensing classification. *Algorithms* **2018**, *11*, 9. [CrossRef]
26. Desai, P.; Pujari, J.; Sujatha, C.; Kamble, A.; Kambli, A. Hybrid Approach for Content-Based Image Retrieval using VGG16 Layered Architecture and SVM: An Application of Deep Learning. *SN Comput. Sci.* **2021**, *2*, 170. [CrossRef]
27. Dolgikh, S. Unsupervised Generative Learning and Native Explanatory Frameworks. *Camb. Open Engag.* **2020**. [CrossRef]
28. Abukmeil, M.; Ferrari, S.; Genovese, A.; Piuri, V. Survey of Unsupervised Generative Models for Exploratory Data Analysis and Representation Learning. *ACM Comput. Surv.* **2021**, *54*, 99. [CrossRef]
29. Xie, J.; Wu, N.Y. *Generative Model and Unsupervised Learning in Computer Vision*; University of California: Los Angeles, CA, USA, 2016. Available online: https://escholarship.org/uc/item/7459n9w5#main (accessed on 4 May 2022).
30. Coates, A.; Ng, A.; Lee, H. An analysis of single-layer networks in unsupervised feature learning. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, Fort Lauderdale, FL, USA, 11–13 April 2011; pp. 215–223.
31. Khan, A.; Sohail, A.; Zahoora, U.; Qureshi, A.S. Asurveyoftherecentarchitecturesofdeepconvolutionalneuralnetworks. *Artif. Intell. Rev.* **2020**, *53*, 5455–5516. [CrossRef]
32. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems. 2012. Available online: https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html (accessed on 4 May 2022).
33. Kaur, S.; Aggarwal, D. Image content based retrieval system using cosine similarity for skin disease images. *Adv. Comput. Sci. Int. J.* **2013**, *2*, 89–95.
34. Tian, Y.; Lei, Y.; Zhang, J.; Wang, J.Z. Padnet: Pan-density crowd counting. *IEEE Trans. Image Processing* **2019**, *29*, 2714–2727. [CrossRef] [PubMed]
35. Pitteri, G.; Munaro, M.; Menegatti, E. Depth-based frontal view generation for pose invariant face recognition with consumer RGB-D sensors. In *International Conference on Intelligent Autonomous Systems*; Springer: Cham, Switzerland, 2016; pp. 925–937.
36. Lu, J.; Behbood, V.; Hao, P.; Zuo, H.; Xue, S.; Zhang, G. Transfer learning using computational intelligence: A survey. *Knowl.-Based Syst.* **2015**, *80*, 14–23. [CrossRef]
37. Li, J.; Wang, J.Z. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. Pattern Anal. Mach. Intell.* **2003**, *25*, 1075–1088.

38. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef] [PubMed]
39. Ermolaev, A.M. Atomic states in the relativistic high-frequency approximation of Kristic-Mittleman. *J. Phys. B At. Mol. Opt. Phys.* **1998**, *31*, L65. [CrossRef]
40. Qureshi, A.S.; Khan, A.; Zameer, A.; Usman, A. Wind power prediction using deep neural network based meta regression and transfer learning. *Appl. Soft Comput.* **2017**, *58*, 742–755. [CrossRef]
41. Wu, S.; Zhong, S.; Liu, Y. Deep residual learning for image steganalysis. *Multimed. Tools Appl.* **2018**, *77*, 10437–10453. [CrossRef]
42. Yuan, Z.W.; Zhang, J. Feature extraction and image retrieval based on AlexNet. In *Eighth International Conference on Digital Image Processing (ICDIP 2016)*; International Society for Optics and Photonics: Bellingham, WA, USA, 2016; Volume 10033, p. 100330E.
43. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
44. Shahriari, A. Visual Scene Understanding by Deep Fisher Discriminant Learning. Ph.D. Thesis, The Australian National University, Canberra, Australia, 2017.
45. Pan, Z.; Yu, W.; Yi, X.; Khan, A.; Yuan, F.; Zheng, Y. Recent progress on generative adversarial networks (GANs): A survey. *IEEE Access* **2019**, *7*, 36322–36333. [CrossRef]
46. Liu, Y.; Zhang, D.; Lu, G.; Ma, W.Y. A survey of content-based image retrieval with high-level semantics. *Pattern Recognit.* **2007**, *40*, 262–282. [CrossRef]
47. Mutasem, K.A. An efficient similarity measure for content based image retrieval using memetic algorithm. *Egypt. J. Basic Appl. Sci.* **2017**, *4*, 112–122. [CrossRef]
48. Kumar, A.; Kumar, B. A review paper: Noise models in digital image processing. *Signal Image Processing Int. J.* **2015**, *6*, 2. [CrossRef]