Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Hybrid Convolutional Neural Network with Domain adaptation for Sketch based Image Retrieval

A. Gheitasi, H. Farsi, S. Mohamadzadeh*

Department of Electrical Engineering, Faculty of Electrical and Computer Engineering, University of Birjand, Birjand, Iran.

Article Info

Article History:
Received 22 March 2024
Reviewed 29 April 2024
Revised 17 June 2024
Accepted 29 June 2024

Keywords:
Sketch Based Image Retrieval (SBIR)
Hybrid CNN
domain adaptation
deep learning

*Corresponding Author's Email Address:
s.mohamadzadeh@birjand.ac.ir

Abstract

Background and Objectives: Freehand sketching is an easy-to-use but effective instrument for computer-human connection. Sketches are highly abstract to the domain gap, that exists between the intended sketch and real image. In addition to appearance information, it is believed that shape information is also very efficient in sketch recognition and retrieval.

Methods: In the realm of machine vision, comprehending Freehand Sketches has grown more crucial due to the widespread use of touchscreen devices. In addition to appearance information, it is believed that shape information is also very efficient in sketch recognition and retrieval. The majority of sketch recognition and retrieval methods utilize appearance information-based tactics. A hybrid network architecture comprising two networks—S-Net (Sketch Network) and A-Net (Appearance Network)—is shown in this article under the heading of hybrid convolution. These subnetworks, in turn, describe appearance and shape information. Conversely, a module known as the Conventional Correlation Analysis (CCA) technique module is utilized to match the range and enhance the sketch retrieval performance to decrease the range gap distance. Finally, sketch retrieval using the hybrid Convolutional Neural Network (CNN) and CCA domain adaptation module is tested using many datasets, including Sketchy, Tu-Berlin, and Flickr-15k. The final experimental results demonstrated that compared to more sophisticated methods, the hybrid CNN and CCA module produced high accuracy and results.

Results: The proposed method has been evaluated in the two fields of image classification and Sketch Based Image Retrieval (SBIR). The proposed hybrid convolution works better than other basic networks. It achieves a classification score of 84.44% for the TU-Berlin dataset and 82.76% for the sketchy dataset. Additionally, in SBIR, the proposed method stands out among methods based on deep learning, outperforming non-deep methods by a significant margin.

Conclusion: This research presented the hybrid convolutional framework, which is based on deep learning for pattern recognition. Compared to the best available methods, hybrid network convolution has increased recognition and retrieval accuracy by around 5%. It is an efficient and thorough method which demonstrated valid results in Sketch-based image classification and retrieval on TU-Berlin, Flickr 15k, and sketchy datasets.

Introduction

Simple Sketches are prevalent in daily life because, Freehand sketching is an easy-to-use but effective instrument for communication and computer-human

connection. Since, touch screens are so commonly used in portable electronics, finding simple Sketches has received more attention. Sketches typically depict location and shape, but actual photographs also include additional useful details like color and texture [1]. Compared to actual images, Sketches are considered as highly sparse signals, and since they are abstract and lack input data [2], their analysis might be difficult. As a result, comparing low-detail images with real photos with a high pixel density is challenging [3]. The following are some of the difficulties in this field:

1. While Sketches are highly abstract and have very little shape information, natural images provide a great deal of color and texture information.

2. On the other hand, as every individual displays an object's sketch uniquely, there will be a wide range of Sketch styles for an object.

3. The training data, may not include every query that might be asked.

4. Another significant obstacle in this sector is the domain gap, that exists between the intended sketch and real image.

5. In each class of the accessible dataset, a small number of Sketches are compared to real images [4]. Strategies for sketch-based image recognition and retrieval are generally similar. Handcrafted features including the Fisher vector, Scale Invariant Feature Transform (SIFT) Strategies for sketch-based image recognition and retrieval are generally similar. Handcrafted features including the Fisher vector, SIFT [5], GF-HOG [6], Histograms of Oriented Gradients (HOG) and Structural Similarity Index Measure (SSIM) were employed in earlier methods. To acquire broad features, these descriptors are frequently paired with a Bag Of Words (BOW) [7]. The main factor in differentiating learning Sketches is their unique and potent features [8]. Until now, the majority of models have solely considered account visual details like color and texture. Yet, shape information has not been the subject of many investigations. Simultaneous consideration of appearance and shape information leads to considerable improvements in sketch-based image recognition and retrieval.

In image processing researches, the use of various features has always been confirmed, taking into account low-level and high-level features at the same time, features such as depth, shape, appearance [9], semantic features, map features [10], visual and time features [11], texture, hierarchy [12], etc. Based on the problem nature, the appearance and shape features are considered at the same time, so that both the general appearance information of the Sketch and the features related to its shape are considered since, the Sketch consists of a series of simple lines. Therefore, the most comprehensive features that can be effective are these two features. Although there have been other studies in the past, that have used these two features in the retrieval problem like [9], the idea of the upcoming research is to use the architecture of Inception V1 in the branch of extracting the appearance feature with the aim of using filters with different dimensions at the same time, which improves the performance. In addition, the use of the domain adapting module, along with this hybrid network, had a positive effect on the retrieval accuracy, which has not been used in similar research.

Therefore, the current study aims to improve the performance of the SBIR system, by only using and analyzing the Sketches, not real images. Accordingly, the best model has been proposed by simultaneously considering the appearance features and shape features through a two-branch network and testing the architectural combination of different networks. In addition, with the aim of reducing the domain distance between Sketches and images, and achieving a more accurate recovery system, the CCA technique module, has been used [13].

Related Works

Since the early 1990s, SBIR was explored, and between 1990 to 1994, Content-Based Image Retrieval (CBIR) was a popular topic, which is still an interesting topic for the researches. Research on SBIR started in 1994 [7]. SBIR is a cross-domain retrieval research where the dataset is made up of real images and the query is run from an abstract sketch [14]. As a result, the majority of studies focus on identifying traits that Sketches and images have in common. For SBIR, training methods extract shallow features [4]. The previously stated SIFT [5], GF-HOG [6], HOG [7], SSIM [15], Soft Histogram of Edge Logal Orientations (SHELO) [16], Fisher vector [17], etc. are the major emphasis of low-level features. Furthermore, several studies also aim to get features at the medium level.

Pixel-based methods often have limited flexibility and high processing costs. Subsequently, a feature extraction module was added to extract several features, with significant edge changes. Sabeti [18], first decomposed the image into blocks of equal size and then extracted a feature vector from each block. Subsequently, a tree-structured hierarchical clustering technique, was developed to divide the blocks into multiple classes based on the extracted features and train a classifier for each class to determine whether a block is from the same overlapping image. In some studies like [19], spectral and spatial features have been used simultaneously for image processing.

In addition, a binary mask was used for objects that matched the original image spatially. The gallery display module also combines the best-bin-first and K-means tree. The amalgamation of these two methods resulted in

a multiplication of the retrieval speed. Moreover, the multiscale representation of edge maps was used to illustrate changes in the degree of detail contained in Sketches created by humans. They found that the intricacies of the sketch were preserved by the mixture of scales [20]. In another research, the K-means fuzzy clustering is used in combination with the convolution neural network [21]. Oriented Chamfer Matching (OCM) is another pixel-based method that examines the sketch's closest edge pixel that matches the image. Sain et al. [22] introduced a novel network that can cultivate certain hierarchies and make use of them to match the Sketch with the image in the appropriate hierarchical levels.

Research in the field of SBIR has recently taken on a new shape with the advent of deep learning and the application of deep neural networks [23]. Due to Deep Convolutional Neural Network's (DCNN's) effectiveness, researchers have also created potent deep learning-based methods that can model intricate sketch aspects. Convolutional networks decrease ambiguities and defects in data and are thorough and effective in image processing [24]. In general, neural network-based methods outperform conventional statistical models and offer stronger data pattern identification, faster processing, and greater resistance to environmental changes [25]. Custom architectures for feature ranking and prediction have been employed recently, including multi-objective ranking networks [26], hybrid CNN models [9], Alex-Net, and Google-Net [27]. In [9], although the combined network of S_net and A_net has used, it has not done anything to reduce the distance of sketch and image nor studied about different network architectures. [28] has extracted the deep features of the sketch using a hybrid CNN model based on ImageNet, but it can be criticized because it did not use a proper pre-processing and only used canny edge extraction. In [29], a deep convolutional three-branch neural network named Sketch Net with the Soft_max function as the cost function is used in a weakly supervised method. Bhattacharjee et al. [30] claims that by proposing an adaptive search method, the query is able to locate small objects in complex background, but only used appearance information to disambiguate between object proposals and refine search results. [31] solved the SBIR problem as a three-stage solution consisting of a self-awareness module for feature learning, a mutual-awareness module to generate probabilistic matches, and a kernel-based communication network to aggregate local matches. Prolonging the process of solving the problem and not considering the pre-processing stage of the data are among the points that have not been analyzed in this research. Chauhan [32] developed a novel method for Zero Shot Sketch Based Image Retrieval (ZS-SBIR) called aligned adaptability and generalizability Adaptability Balance Domain and Generalizability (ABDG). Generally speaking, this method aligns discriminability with generalizability learning for each instance by applying a two-stage advanced knowledge sketch. An end-to-end deep network termed DVSF is presented in [33] for sketch recognition. It feeds deep data to visual and sequential networks at the same time to acquire temporal and spatial features. Some researchers have proposed and solved the problem in a fuzzy way [34], [35]. A deep structure known as Network In Network (NIN)—a micro-neural network with a complicated structure—was shown in research [36]. A robust performance estimator, the multilayer perceptron, is used in the creation of this network. Wang et al.'s network with contrast cost function [37] was one of the first studies on multi-branch networks for sketch retrieval (3D objects). Each branch's weights were individually trained to correspond with the sketch and image domains. Some other examples of research that have studied multi-branch and hybrid networks in the subject of image retrieval are introduced below. Visual Geometry Group (VGG) Net is a CNN-based network that uses data augmentation techniques in conjunction with dimensionality reduction of the output layer [38]. Qi et al. [39] recently suggested a bifurcated Siamese network with contrast cost function. A triple CNN—an anchor branch—is used by Tu et al. [40] as a reference object. The use of triple CNNs for face recognition [41], tracking, visual search for photography, and queries intended to refine search inside a class [3] have recently been investigated. Similarly, Sangkloy et al. [42] used the Sketchy system to implement a Fine-Grained SBIR method.

Additionally, a multi-stage, triple network based on CNN is suggested in [43]. A straight forward and effective method that learns the semantic embedding space from a vision model and does not require a lot of computer training resources is shown in [44]. Indeed, a pre-trained Image Net CNN is used here along with three cost functions: semantic knowledge retention, domain adapter, and semantic classifier. According to the experts, utilizing a completely shared network is preferable to utilizing a bifurcated network without weight sharing. On the other hand, the authors of [42] argued that considering cross-category retrieval, it is better to avoid layer sharing. Additionally, Bui et al. [45] analyzed a hybrid sketch that shared certain layers but had identical architecture in both branches.

Experiments conducted on actual datasets demonstrate that deep features outperform handmade descriptors in SBIR. A sketch is first transformed into a collection of points, and then CNNs are used to describe the shape information based on the set of points to acquire strong shape features. Finally, CNN's two-stream Sketch combines appearance and shape information to provide unique features. To lessen the domain gap, the domain adaptation module is also utilized under the title of the conventional correlation analysis approach module CCA [13].

Proposed Method

This section presents the bifurcated hybrid convolutional network's architecture and discusses the specific details of each component.

A. Hybrid Convolutional Network Architecture

Generally, shape and appearance data may be used to characterize a sketch. To integrate appearance and shape information, a bifurcated hybrid convolutional network is therefore suggested. Fig. 1 depicts this network's overall structure. The definition of the A-Net network, the first branch, describes appearance data, whereas the S-Net network, the second branch, describes and extracts shape features.

The two networks have distinct inputs. The A-Net network's input directly determines the main sketch image. The probabilities of each object class constitute the branch's output, and following feature extraction, the feature vectors are sent to all linked layers.

On the other hand, many points are initially sampled from the sketch image to acquire the shape's features. The feature vectors related to the shape are then extracted using layers like sketch layers, alignment layers, perceptron, and pooling. S-Net specifications are covered in more depth later.

We have both appearance features and shape features when A-Net and S-Net are trained separately. Subsequently, the two feature vectors undergo independent concatenation, resulting in a composite feature representation. In other words, appearance and shape feature vectors are obtained from A-Net and S-Net branches, respectively. These vectors are normalized independently and separately, and then we put them together and use them as a combined feature vector for sketch-based image retrieval and classification. Finally, pattern classification and pattern-based image retrieval are achieved through the utilization of the hybrid features. The hybrid features are fed into a Support Vector Machine (SVM) classifier for sketch classification.

The studied sketch features and the resulting edge maps are taken from the image dataset for vector SBIR. The distance is then compared to determine the ranking results. The tests' findings demonstrate how image classification and retrieval are enhanced when hybrid appearance and shape data are used.

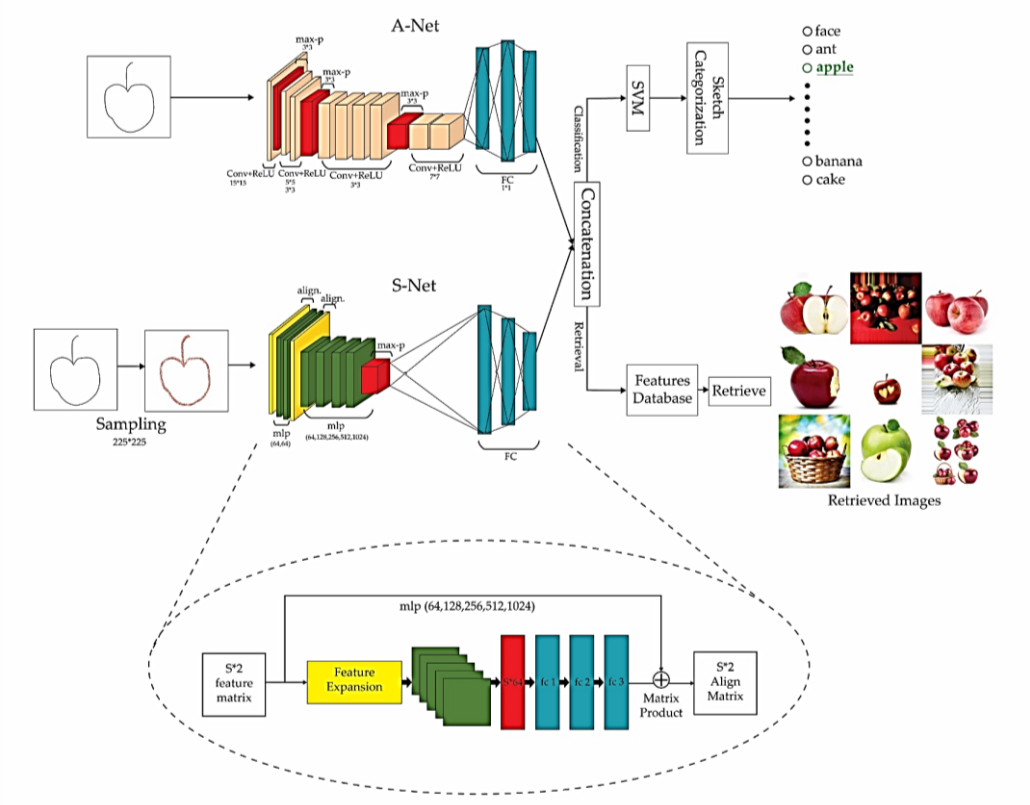


Fig. 1: The general framework of the proposed bifurcated hybrid convolutional network.

To show the shape information for the S-Net network, the sketch is first transformed into a set of points, from which features are then retrieved. Compared to other descriptors, the shape features represented by the point set are distinct. There are some problems in this regard: The finished drawings have distinct styles when they are done by various artists. For various wheels and Sketch styles, the learned representation of sample points should be uniform. As mentioned above, a Sketch is described by extracting a large number of points. Various

starting points for sampling might result in different perturbations for the same Sketch.

S-Net architecture is suggested as a solution for these problems. The class label is the output of the S-Net network, which receives the collection of points as input. A pre-processing step is necessary to create the set of points that make up the S-Net input from the sketch. Sampling with 512 points, has been chosen based on testing this sampling method with a variety of points and evaluating the network's performance, to minimize the computation's time complexity.

The set of collected points is aligned with a conventional space after acquiring the sampled point image. Subsequently, an MLP network extracts the features. Alternatively, the set of points is aligned in the feature space using a transformation matrix. The sketch's high-level shape features are extracted using a Multi-Layer Perceptron (MLP) network that has fully linked layers and a pooling layer. In the below the general Algorithm of the proposed method has been shown.

Algorithm 1: Algorithm of the proposed method

INPUT:

Train Img Set: training set that is images are randomly selected in 9 set 8, 16, 24, 32, 40, 48, 56 and 72 of Sketches
Test Img Set: randomly selected of 10% of the remaining data
Pre-Train: $\mathbf{I_A}$: All input Sketches are resized to 225*225
$\mathbf{I_S}$: 512- point Farthest Point Sampling (FPS) & Resized to 225*225

1. Normalization of the training set for classification and retrieval
2. For a query sketch $\mathbf{I_A}$
   Concatenate the deep visual and extract appearance features
   Return the probability of each of each object class
3. For a query sketch $\mathbf{I_S}$
   Concatenate the alignment to predict a M = affine transformation matrix and extract shape features
   Return the label of class
4. Training step:
   For e = 1:E do E = number of epochs
   a) Feed the sketch in to A-Net & S-Net
      $\mathcal{F}_{fc_2}^{app}(\boldsymbol{I_A}) \leftarrow \boldsymbol{I_A}$ /*sketch appearance feature mapping*/
      $\mathcal{F}_{fc_2}^{shape}(\boldsymbol{I_s}) \leftarrow \boldsymbol{I_s}$ /*sketch shape feature mapping*/
   b) Calculate the ranking loss according to (1) and update the model parameters
   c) $\mathcal{F}^{+} \longleftarrow \left[\mathcal{F}_{fc_2}^{app}, \mathcal{F}_{fc_2}^{shape}\right]$ combining feature vectors

End

5. Ranking:
   a) Classification: using SVM classification
   b) SBIR: - using CCA convolutional correlation analysis technique module
      - using K-Nearest Neighbors (KNN) to find the closest answer to the query sketch between the sketch query and top-k samples in testing set

OUTPUT:

Some of the closest real images to the query sketch

1. Preprocessing

The images are initially resized to 225 × 225 during the pre-processing phase. The background is represented by one and lines by zero. Then, as X $= \{x_1, x_2, \ldots, x_n\}$, a subset of points taken from the sketch are chosen. Although the set of points might be selected at random, the findings indicate that it is preferable to have a minimum amount of space between the points. As a result, the Farthest Point Sampling (FPS) sub-set of points' furthest repetitive point is chosen using sampling.

Compared to random sampling, using this sampler offers greater coverage for a sketch and as close to uniform sampling as possible. As previously stated, 512 points are sampled for each sketch, and the coordinates (x,y) of each point are used to signify it.

2. Alignment Stage

Rotation and transformation/conversion are two examples of geometric transformations in which the semantic information of the sampled points of the Sketches should be invariant. To align the input space, an alignment network is built to predict a two-by-two affine transformation matrix. The set of input points is then multiplied directly in the anticipated transformation matrix. The whole input alignment procedure is depicted in Fig. 1.

The alignment phase is carried out twice in the suggested network architecture: once in the input space and once in the feature space. Although the transformation matrix in the feature space is 64x64, the alignment in these two spaces is similar. Three completely

linked layers with output sizes of 1024, 512, and 256 are inserted in the continuation of Max Pooling. The output sizes of each layer are 64, 128, 256, 512, and 1024, respectively.

A regularization component is introduced to the cross entropy as a function of training cost because optimization becomes challenging when the feature space's transformation matrix has much larger dimensions than the original input space.

$$\ell = \mathcal{T}(\theta) + \Im \times \mathcal{L} \tag{1}$$

$$\mathcal{T}(\theta) = \frac{-1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\left(y^{(i)} = j\right) log\frac{exp(a_k)}{\sum_{k^*=1}^{c} exp(a_{k^*})}\right] \tag{2}$$

$$\mathcal{L} = \|I - MM^T\|_2^2 \tag{3}$$

$$a_k = \sum_j W_{kj} Z_j \tag{4}$$

where $\Im$ is the $\mathcal{L}$ function's alignment weight. Also, it was found that $\Im = 0.01$ provides the best accuracy based on the testing and errors. The output of hidden unit j is Z_j. The feature alignment matrix is M, and the input is I [9]. By including regularization into the cost function, the orthogonal transformation preserves the original data, resulting in stable optimization and improved model performance. Three input layers, one hidden layer, and one output layer exist in every MLP. In the suggested MLP, five hidden layers—64, 128, 256, 512, and 1024 neurons—are added to the MLP after the second alignment, and two hidden layers—64 and 64 neurons—are added to the S-Net network following the first alignment. The points' irregularity is crucial for the shape descriptor.

Max Pooling is used to collect data from every point to ensure the consistency of the collection of sampled points. In this case, n vectors that were computed by the preceding MLP module are the input that Max Pooling uses. As a result, a fresh vector that supplies the completely linked layers is produced.

Fully linked layers apply an offset bias and multiply their input by a weight matrix. The S-Net network consists of three completely linked layers, with the number of Sketch categories being the output of the third layer.

B. Hybrid Convolutional Network Training

The appearance features and shape features are combined using a hybrid procedure that joins $\mathcal{F}_{fc_2}^{app}$ and $\mathcal{F}_{fc_2}^{shape}$ shape with ℓ_1 normalization specified by $\mathcal{F}^+$ after the A-Net and S-Net have been trained separately [9].

$$\mathcal{F}^+ \leftarrow \left[\mathcal{F}_{fc_2}^{app}, \mathcal{F}_{fc_2}^{shape}\right] \tag{5}$$

A binary SVM is separately trained for every category to classify the sketch. Prediction scores are produced for each category at the time of evaluation, which makes direct use of the integrated features of each design. The next section will provide the features of the sketch classification.

The edge maps of the natural image data set are initially extracted using gpb to reconstruct the image-based plan.

Subsequently, features are extracted on the resultant edge maps and the sketch in question separately using the hybrid CNN. Finally, K top candidate images for the given sketch are retrieved using K-Nearest Neighbors (KNN) classification with Euclidean distance. The hybrid CNN network's retrieval performance is covered in the evaluation section.

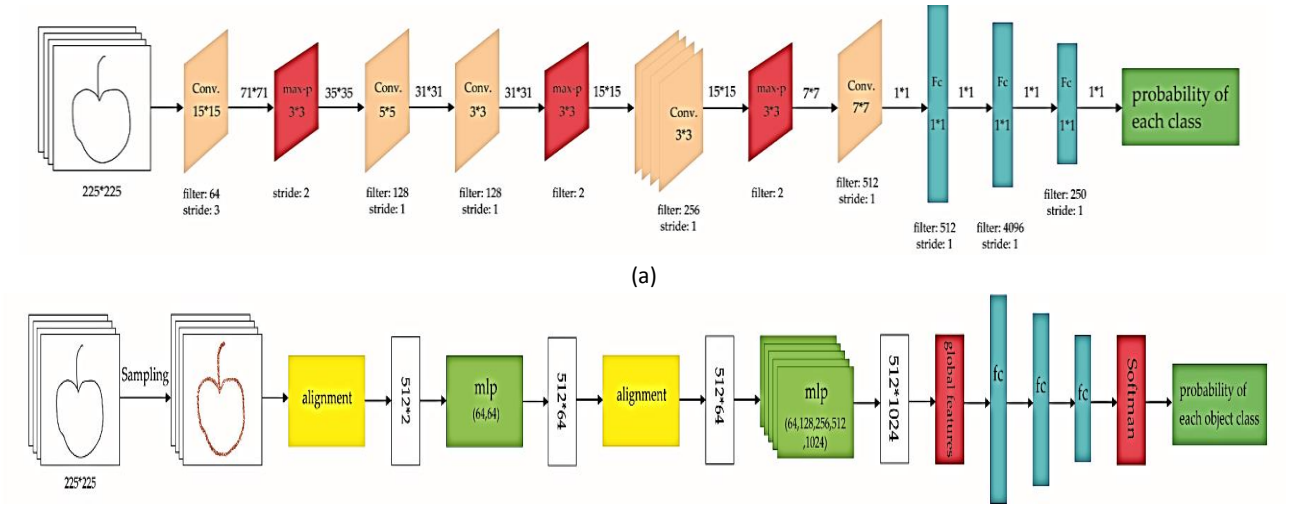


Fig. 2: Architectural details of (a) A-net and (b) S-net.

Fig. 2 shows the architecture details of two branches of the proposed network. As shown, a CNN network consists of convolution and max pooling layers, which end up with fully connected layers. Testing various examples of CNN, including VGG, Google Net, Alex-Net, Sketch A Net and Inception V1 for A-Net branch in combination with alignment network (S-Net) demonstrated that the CNN network with the Inception V1 module has the best performance.

Results and Discussion

This section will analyze the suggested hybrid CNN network. First, a description of the evaluation's data set is given. Then, the network's performance and its results are discussed.

A. Evaluation Criteria

1. Accuracy precision

One of the most frequently used evaluation criteria in classification problems is accuracy. It is determined by dividing the total number of identified samples (both properly and wrongly categorized) by the ratio of correctly classified samples [16]. The following formula is used to calculate accuracy.

$$Precision = \frac{TP}{TP + FP} \times 100 \quad (6)$$

where FP and TP represent samples that were incorrectly identified and those that were correctly identified.

2. Recall

The following formula may be used to represent recall, which is a metric derived from the ratio of properly classified samples to the total of correctly identified and incorrectly rejected samples [46].

$$Precision = \frac{TP}{TP + FP} \times 100 \quad (7)$$

3. MAP (Mean Average Accuracy)

The weighted mean of the accuracy at each threshold is used to compute the mean accuracy. Recall gains weight over the preceding level. The mean AP for every class determines the mean accuracy [47].

$$AP = \frac{1}{N} \sum_{r} P_{interp}(r) \quad (8)$$

where MAP is the mean AP in each data class and $P_{interp}(r)$ is the precision in calling point(r).

B. Datasets

The two datasets used for classification evaluation are Sketchy and TU-Berlin.

1. TU-Berlin

This collection of 20,000 hand-drawn images, created by regular people, is divided into 250 groups, the majority of which include common everyday items like eyes, aircraft, apples, and bananas. Each design measures 1111×1111. There are eighty sketch images collected for every class.

2. Sketchy

There are 125 classifications in the collection, and there are 75471 human Sketches for items. In this study, the remaining Sketches were utilized for training, and 50 Sketches from each category were used for testing.

3. Flickr 15k

15024 images make up this data collection, from which samples of Sketches have been taken. Ultimately, 33 categories have been created from the labeled sketch images, with 10 Sketches in each category. Using data augmentation methods, this data set has been augmented during the review process. Among the methods employed for data augmentation were: 1. Vertical and horizontal flip 2. Turning angles of ±10 and ±15 degrees 3. Magnification ±1.2 and ±1.5 times the design height; 4. Taking into account the planar movement of Sketches, translation in both horizontal and vertical directions with ±10 and ±15 pixels is taken into account.

C. Training and Testing

TensorFlow's Stochastic Gradient Descent (SGD) is used to train the model. The batch size is 64, the momentum is 0.9, and the learning rate is 0.01 for the A-Net network. Every period of the optimization procedure involves a 0.3 reduction, and it is terminated after 15 optimization periods. The initiating weights of the bias terms have a Gaussian distribution with δ=0.01 and μ=0, and they are initialized from 0.1. The final three completely linked layers of the S-Net network employ elimination at a rate of 0.7. For batch normalization, the decay rate initiates at 0.5 and rises steadily to 0.99. Additionally, an Adam optimizer with a batch size of 64, a momentum of 0.9, and a learning rate of 0.001 was utilized. The learning rate in each of the five courses is split in two.

Nine educational division models are put to the test. For every class, a random selection of 8, 16, 24, 32, 40, 48, 56, and 72 Sketches is used as the training data set. To estimate robust model parameters, however, 10% of the remaining data are chosen at random to serve as the validation data set.

D. Evaluation

Table 1 and Table 2 show that the following results can be attained:

- Compared to other basic networks, the proposed hybrid convolution performs better. As shown, it achieves a classification score of 84.44% for the TU-Berlin dataset and 82.76% for Sketchy.
- Tables 1 and 2's results indicate that NIN performs better than VGG-Net when there are less than 50

training plans, while VGG-Net performs better than NIN when there are more than 50 training plans. However, as compared to base networks, the hybrid CNN performs better across all training set sizes.

- Deep features outperform hand-made features in terms of performance; nonetheless, the more images in the training plan, the better the performance to the extent that the classification accuracy is higher than human accuracy, even with a significant amount of training images. With 45 training Sketches, the system achieves superhuman accuracy on average for each category. However, a hybrid convolutional network needs just 35 training samples to attain this limit.

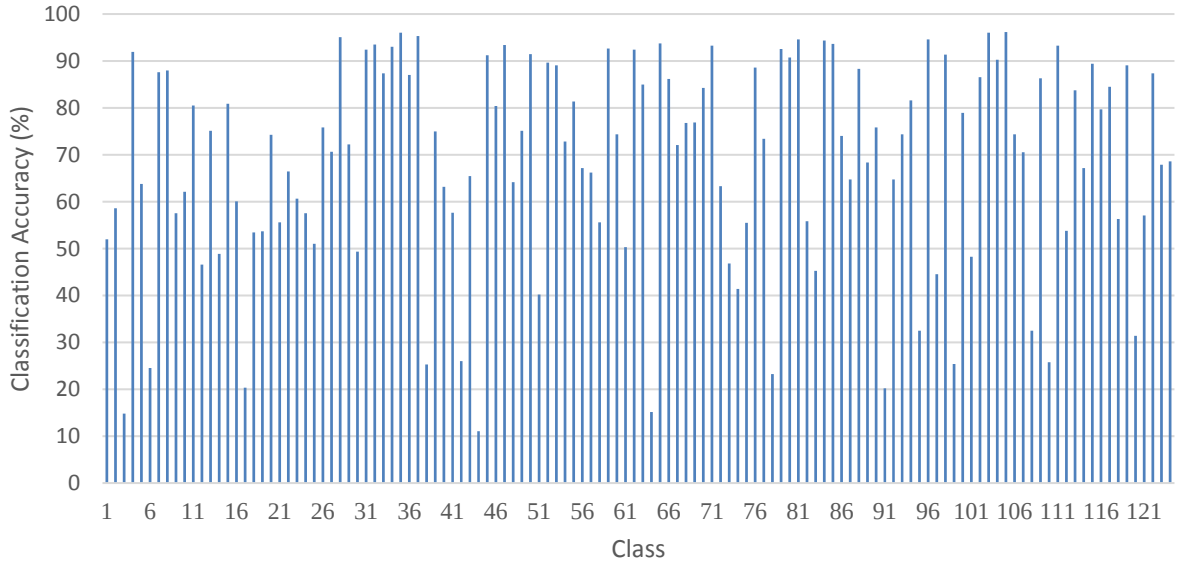
Table 1: Classification accuracy of different features on the TU-Berlin dataset

Method	Size of training set						
	10	**20**	**30**	**40**	**50**	**60**	**70**
DVSF [33]	60.52	67.43	71.85	73.96	75.87	77.03	78.94
Sketch Net [29]	59.91	65.82	70.06	73.02	75.41	76.25	78.00
VGG Net [38]	51.22	61.33	66.43	66.96	70.09	74.26	86.02
NIN [36]	54.13	64.42	67.34	70.74	71.87	73.00	74.86
Fisher Vector Size 24 [17]	44.96	54.21	58.32	62.18	65.14	66.58	67.86
Fisher Vector Size 16 [17]	42.56	50.37	54.17	58.40	60.85	63.64	65.03
Eitz (SVM soft) [48]	35.00	42.20	45.63	50.12	51.11	54.00	55.20
Eitz (SVM hard) [48]	33.10	39.42	44.16	48.04	49.14	52.13	53.00
SSIM [15]	28.97	38.00	43.52	46.11	48.35	50.91	52.01
Proposed Method	**61.81**	**67.91**	**73.00**	**75.21**	**76.14**	**80.20**	**84.44**

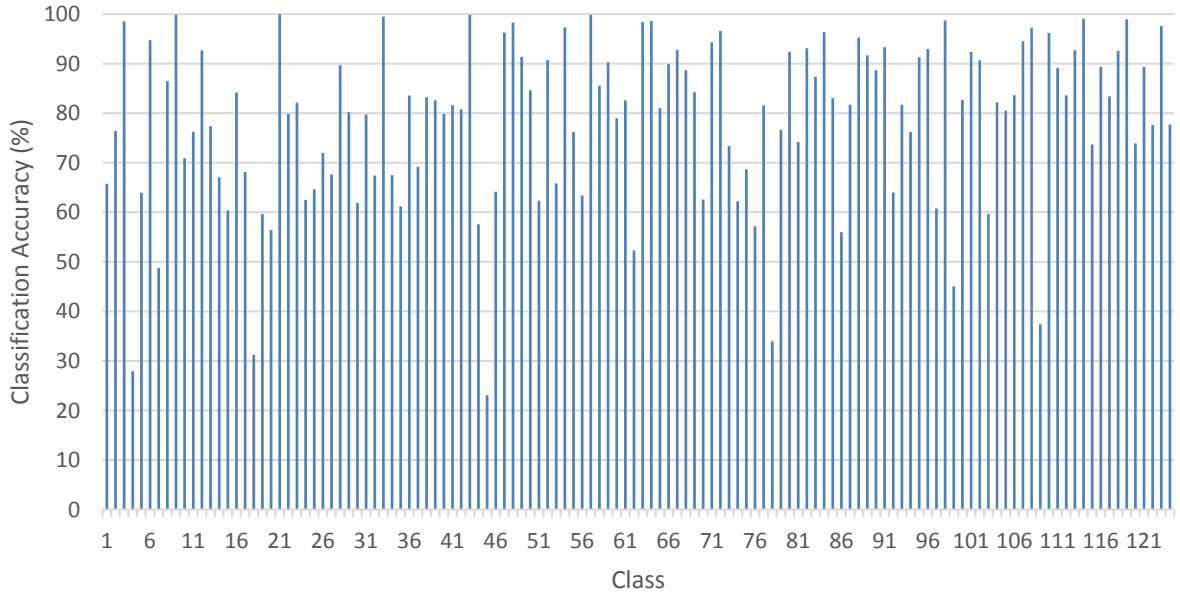
Table 2: Classification accuracy of different features on Sketchy dataset

Method	Size of training set						
	10	**20**	**30**	**40**	**50**	**60**	**70**
DVSF [33]	58.03	64.05	69.21	72.76	74.00	75.18	77.58
Sketch Net [29]	57.53	63.41	68.92	71.94	73.26	74.88	76.84
VGG Net [38]	51.28	59.98	62.27	66.54	70.11	72.85	74.12
NIN [36]	52.02	61.51	65.44	68.07	69.16	70.19	71.96
Fisher Vector Size 24 [17]	43.48	51.27	55.05	59.09	62.20	64.38	66.45
Fisher Vector Size 16 [17]	42.89	50.23	54.08	58.98	61.03	63.67	65.13
Eitz (SVM soft) [48]	32.24	38.76	43.37	45.88	49.14	51.96	53.12
Eitz (SVM hard) [48]	30.24	37.07	42.16	46.14	47.10	48.29	51.94
SSIM [15]	26.92	35.61	40.66	43.07	45.18	46.71	48.96
Proposed Method	**59.98**	**65.78**	**71.16**	**73.26**	**74.97**	**77.53**	**82.76**

Using the Sketchy dataset, the classification performance is evaluated with and without taking the shape feature into account. Fig. 3 illustrates how, when shape features are included, the classification accuracy of seven categories is less than 45% and the accuracy of more than 50% of categories is above 70%.



(a)



(b)

Fig. 3: Classification performance of the hybrid CNN on Sketchy dataset a) shape feature-free b) including shape feature.

As shown in Fig. 3, the performance of the S-Net branch is lower than the A-Net branch. In addition, considering shape information, the proposed combined network has performed better than the A-Net branch alone. These results indicated that the extraction and use of shape features can enhance the classification performance of the scheme.

Another experiment was conducted to investigate the impact of different network architectures and choosing the best architecture for the proposed network. The results of this experiment are shown in Fig. 4.

The results of Fig. 4, indicated that the convolution with Inception V1 module has performed better than other architectures in the field of classification. After Inception V1, Alex-net and Google-net are the architectures that can be used in the field of image retrieval with a slight difference.

Table 3 compares the results of several methods using the Sketchy dataset and the MAP criterion.

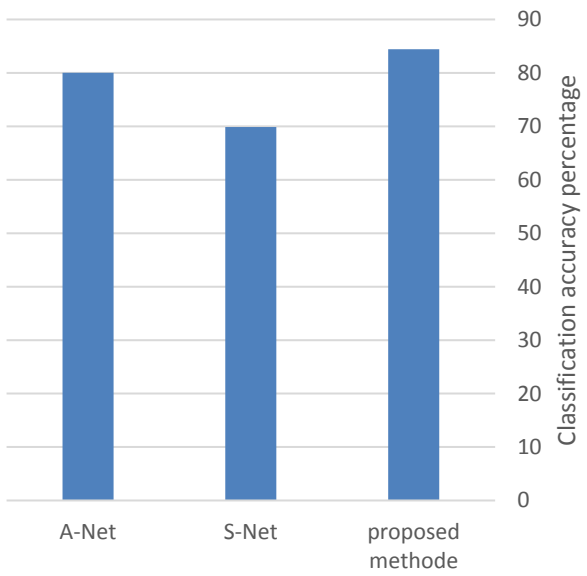


Fig. 4: The effect of each branch of the proposed network alone and combined.

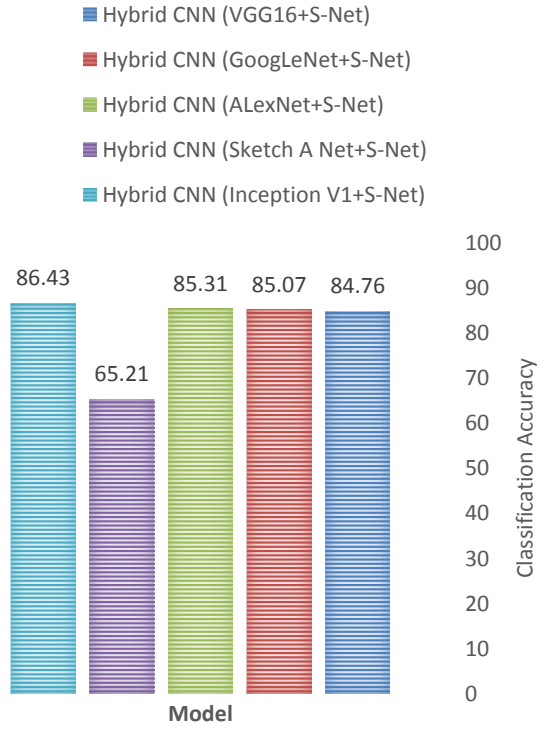


Fig. 5: Testing different architectures in combination with S-Net network.

The SBIR is tested using the expanded 15k Flickr dataset; for this evaluation, 2/3 of the dataset is used for training and 1/3 for testing. The evaluation's findings on the expanded Flickr 15k dataset are displayed in Table 4.

It is evident from the table data that all deep learning-based methods outperform hand-made feature extraction methods.

Table 3: Results of SBIR based on the MAP criterion on the Sketchy dataset

Method	MAP
HOG [49]	0.115
GF-HOG [6]	0.157
SHELO [50]	0.161
Siamese With Contrastive Loss [39]	0.195
RST-SP-SHELO [16]	0.200
Sketch A Net [26]	0.208
Triplet Sketch-Edgemap [45]	0.244
Query adaptive Reranking CNN [30]	0.323
Sketchy Triplet [42]	0.359
Sket RET [32]	0.437
Siamese CNN [39]	0.481
Cross Modal [22]	0.523
GN Triplet [42]	0.529
SBT-Net [44]	0.553
CNN with Multi Stage Regression [43]	0.565
Hybrid CNN [9]	0.574
Proposed Method	**0.585**

Table 4: The MAP criterion-based SBIR results on the extended Flickr 15k dataset

Method	MAP
Sketch A Net [26]	0.084
SIFT [51]	0.101
HOG [49]	0.119
Query adaptive Reranking CNN [30]	0.130
GF-HOG [6]	0.132
Sketchy Triplet [42]	0.145
Sket RET [32]	0.177
Siamese CNN [39]	0.195
Cross Modal [22]	0.212
GN Triplet [42]	0.214
SBT-Net [44]	0.224
CNN with Multi Stage Regression [43]	0.229
Triplet CNN [40]	0.245
Hybrid CNN [9]	0.287
Proposed Method	**0.293**

Conclusion

This research presented the hybrid convolutional framework, which is based on deep learning, for sketch retrieval. In addition to appearance information, it is thought that shape information is also very efficient in sketch recognition and retrieval. A hybrid network architecture comprising two networks—S-Net and A-Net—is shown in this article under the heading of hybrid convolution.

These subnetworks, in turn, describe appearance and shape information.

The classification accuracy of the proposed method reached 84.44% on the comprehensive

TU-Berlin dataset and 82.76% accuracy on the challenging Sketchy dataset, which is a significant value. It was also shown how the operation of two branches of the network simultaneously can be effective in the image retrieval process than when each branch is used alone, and this highlights the importance of using appearance and shape features at the same time. On the other hand, using the CCA module to reduce the domain distance of the sketch and image, as well as choosing the appropriate network architecture, improved the efficiency of the proposed retrieval system. Compared to the best available methods, hybrid network convolution has increased recognition and retrieval accuracy by around 5%. It is an efficient and thorough method which demonstrated valid results in sketch-based image classification and retrieval on TU-Berlin, Flickr 15k wide, and sketchy datasets.

Author Contributions

A. Gheitasi, H. Farsi, and S. mohammadzadeh designed the experiments. A. Ghatasi collected the data. A. Ghatasi carried out the data analysis. A. Gheiaasi, H. Farsi, and S. Mohamadzadeh interpreted the results and A. Mahboobi wrote the manuscript. H. Farsi and S. Mohamadzadeh Corrected editing errors.

Acknowledgment

I am sincerely grateful for the guidance and help of honorable professors Mr. Hassan Farsi and Sajjad Mohammadzadeh. This work is completely self-supporting, thereby no any financial agency's role is available.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

SBIR	Sketch Base Image Retrieval
S_Net	Shape Network
A_Net	Appearance Network
CCA	Conventional Correlation Analysis
CNN	Conventional Neural Network
TU_Berlin	Technische Universität Berlin
SIFT	Scale Invariant Feature Transform
HOG	Histograms of Oriented Gradients
SSIM	Structural Similarity Index Measure
BOW	Bag Of Words
OCM	Oriented Chamfer Matching
DCNN	Deep Conventional Neural Network
ZS_SBIR	Zero Shot Sketch-Based Image Retrieval
ABDG	Adaptability Balance Domain and Generalizability
DVSF	Deep Visual-Sequential Fusion
NIN	Network In Network
VGG	Visual Geometry Group
SVM	Support Vector Machine
MLP	Multi-Layer Perceptron
Img	Image
FPS	Farthest Point Sampling
KNN	K-Nearest Neighbors

MAP	Mean average Accuracy Percent
SHELO	Soft Histogram of Edge Logal Orientations
SGD	Stochastic Gradient Descent
GN	Google Net

References

[1] D. Birari, D. Hiran, V. Narawade, "Survey on sketch based image and data retrieval," in Proc. 2nd International Conference on Communications and Cyber Physical Engineering (ICCCE 2019): 285-290, 2020.

[2] A. Chaudhuri, A. K. Bhunia, Y. Z. Song, A. Dutta, "Data-free sketch-based image retrieval," arXiv preprint arXiv:2303.07775, 2023.

[3] A. Sain, A. K. Bhunia, Y. Yang, T. Xiang, Y. Z. Song, "Stylemeup: Towards style-agnostic sketch-based image retrieval," in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[4] P. Xu, T. M. Hospedales, Q. Yin, Y. Zh. Song, T. Xiang, L. Wang, "Deep learning for free-hand sketch: A survey," IEEE Trans. Pattern Anal. Mach. Intell., 45(1): 285-312, 2023.

[5] Z. Hossein-Nejad, H. Agahi, A. Mahmoodzadeh, "Remote sensing image registration based on a geometrical model matching," J. Inf. Syst. Telecommun. (JIST), 5(36): 41, 2021.

[6] R. Hu, S. James, T. Wang, J. Collomosse, "Markov random fields for sketch based video retrieval," in Proc. the 3rd ACM conference on International Conference on Multimedia Retrieval (ICMR): 279-286, 2013.

[7] Y. Li, W. Li, "A survey of sketch-based image retrieval," Mach. Vision Appl., 29(7): 1083-1100, 2018.

[8] S. Mohamadzadeh, S. Pasban, J. Zeraatkar-Moghadam, A. K. Shafiei, "Parkinson's disease detection by using feature selection and sparse representation," J. Med. Biol. Eng., 41(4): 412-421, 2021.

[9] X. Zhang, Y. Huang, Q. Zou, Y. Pei, R. Zhang, S. Wang, "A hybrid convolutional neural network for sketch recognition," Pattern Recognit. Lett., 130: 73-82, 2020.

[10] P. Xu, T. M. Hospedales, Q. Yin, Y. Zh. Song, T. Xiang, L. Wang, "Deep learning for free-hand sketch: A survey," IEEE Trans. Pattern Anal. Mach. Intell., 45(1): 285-312, 2020.

[11] P. Xu, Y. Huang, T. Yuan, K. Pang, Y. Zh. Song, T. Xiang, T. M. Hospedales, Zh. Ma, J. Guo, "Sketchmate: Deep hashing for million-scale human sketch retrieval," in Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018): 8090-8098, 2018.

[12] A. Dutta, Z. Akata, "Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval," in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019): 5089-5098, 2019.

[13] H. Yu, M. Huang, J. J. Zhang, "Domain adaptation problem in sketch based image retrieval," ACM Trans. Multimidia Comput. Commun. Appl., 19(3): 1-17, 2022.

[14] A. K. Bhunia, A. Sain, P. H. Shah, A. Gupta, P. N. Chowdhury, T. Xiang, Y. Zh. Song, "Adaptive fine-grained sketch-based image retrieval," in Proc. 17th European Conference Computer Vision (ECCV 2022), Part XXXVII: 163-181, 2022.

[15] E. Shechtman, M. Irani, "Matching local self-similarities across images and videos," in Proc. 2007 IEEE Conference on Computer Vision and Pattern Recognition: 1-8, 2007.

[16] J. M. Saavedra, "Rst-shelo: Sketch-based image retrieval using sketch tokens and square root normalization," Multimedia Tools Appl., 76(1): 931-951, 2017.

[17] R. G. Schneider, T. Tuytelaars, "Sketch classification and classification-driven analysis using fisher vectors," ACM Trans. Graphics (TOG), 33(6): 1-9, 2014.

[18] V. Sabeti, "An improved approach to blind image steganalysis using an overlapping blocks idea," J. Electr. Comput. Eng. Innovations, 11(2): 263-276, 2023.

[19] M. Imani, "Target detection using multispectral images, A case study: Wheat detection in Chenaran County in Iran," J. Electr. Comput. Eng. Innovations, 9(1): 11-24, 2020.

[20] M. Rezaei, M. Rezaei, "Foreground-back ground segmentation using k-means clustering algorithm and support vector machine," J. Inf. Syst. Telecommun. (JIST), 1(41): 65, 2023.

[21] S. Fooladi, H. Farsi, S. Mohamadzadeh, "Segmenting the lesion area of brain tumor using convolutional neural networks and fuzzy K-means clustering," Int. J. Eng., 36(8): 1556-1568, 2023.

[22] A. Sain, A. K. Bhunia, Y. Yang, T. Xiang, Y. Zh. Song, "Cross-modal hierarchical modelling for fine-grained sketch based image retrieval," arXiv preprint arXiv:2007.15103, 2020.

[23] A. K. Bhunia, P. N. Chowdhury, A. Sain, Y. Yang, T. Xiang, Y. Zh. Song, "More photos are all you need: Semi-supervised learning for fine-grained sketch based image retrieval," in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 4247-4256, 2021.

[24] A. K. Bhunia, S. Koley, A. F. U. R. Khilji, A. Sain, P. N. Chowdhury, T. Xiang, Y. Zh. Song, "Sketching without worrying: Noise-tolerant sketch-based image retrieval," in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019): 999-1008, 2022.

[25] A. Gheitasi, H. Farsi, S. Mohamadzadeh, "Estimation of hand skeletal postures by using deep convolutional neural networks," Int. J. Eng., 33(4): 552-559, 2020.

[26] Q. Yu, F. Liu, Y. Zh. Song, T. Xiang, T. M. Hospedales, Ch. Ch. Loy, "Sketch me that shoe," in Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016): 799-807, 2016.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, "Going deeper with convolutions," in Proc. the IEEE Conference on Computer Vision and Pattern Recognition: 1-9, 2015.

[28] N. Kumar, R. Ahmed, V. B. Honnakasturi, S. Sowmya Kamath, V. Mayya, "Sketch-based image retrieval using convolutional neural networks based on feature adaptation and relevance feedback," in Proc. International Conference on Emerging Applications of Information Technology: 103-113, 2022.

[29] H. Zhang, S. Liu, Ch. Zhang, W. Ren, R. Wang, X. Cao, "Sketchnet: Sketch classification with web images," in Proc. the IEEE Conference on Computer Vision and Pattern Recognition: 1105-1113, 2016.

[30] S. D. Bhattacharjee, J. Yuan, W. Hong, X, Ruan, "Query adaptive instance search using object sketches," in Proc. the 24th ACM International Conference on Multimedia: 1306-1315, 2016.

[31] F. Lin, M. Li, D. Li, T. Hospedales, Y. Zh. Song, Y. Qi, "Zero-shot everything sketch-based image retrieval," in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023): 23349-23358, 2023.

[32] R. Chavhan, Zero-Shot Sketch Based Image Retrieval, Indian Institute of Technology Bombay, 2021.

[33] J. Y. He, X. Wu, Y. G. Jiang, B. Zhao, Q. Peng, "Sketch recognition with deep visual-sequential fusion model," in Proc. the 25th ACM International Conference on Multimedia: 448-456, 2017.

[34] H. Zhao, M. Liu, M. Li, "Feature fusion and metric learning network for zero-shot sketch-based image retrieval," Entropy, 25(3): 502, 2023.

[35] E. Askari, S. Motamed, "Computational model for image processing in the minds of people with visual agnosia using fuzzy cognitive map," J. Inf. Syst. Telecommun. (JIST), 2(42): 102, 2023.

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

[37] F. Wang, L. Kang, Y. Li, "Sketch-based 3d shape retrieval using convolutional neural networks," in Proc. the IEEE Conference on Computer Vision and Pattern Recognition: 1875-1883, 2015.

[38] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," arXiv preprint arXiv:1405.3531, 2014.

[39] Y. Qi, Y. Zh. Song, H. Zhang, J. Liu, "Sketch-based image retrieval via siamese convolutional neural network," in Proc. 2016 IEEE International Conference on Image Processing (ICIP): 2460-2464, 2016.

[40] T. Bui, L. Ribeiro, M. Ponti, J. Collomosse, "Generalisation and sharing in triplet convnets for sketch based visual search," arXiv preprint arXiv:1611.05301, 2016.

[41] M. Rohani, H. Farsi, S. Mohamadzadeh, "Deep multi-task convolutional neural networks for efficient classification of face attributes," Int. J. Eng., 36(11): 2102-2111, 2023.

[42] P. Sangkloy, N. Burnell, C. Ham, J. Hays, "The sketchy database: learning to retrieve badly drawn bunnies," ACM Trans. Graphics (TOG), 35(4): 1-12, 2016.

[43] T. Bui, L. Ribiro, M. Ponti, J. Collomosse, "Sketching out the details: Sketch-based image retrieval using convolutional neural networks with multi-stage regression," Computers & Graphics, 71: 77-87, 2018.

[44] O. Tursun, S. Denman, S. Sridharan, E. Goan, C. Fookes, "An efficient framework for zero-shot sketch-based image retrieval," Pattern Recognit., 126: 108528, 2022.

[45] T. Bui, L. Ribeiro, M. Ponti, J. Collomosse, "Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network," Comput. Vision Image Understanding, 164: 27-37, 2017.

[46] C. Bai, J. Chen, Q. Ma, P. Hao, Sh. Chen, "Cross-domain representation learning by domain-migration generative adversarial network for sketch based image retrieval," J. Visual Commun. Image Represent., 71: 102835, 2020.

[47] A. P. R. G. G. Rajput, "Sketch based image retrieval in large databases using edge features," Int. J. Recent Technol. Eng. (IJRTE), 08: 2277-3878, 2020.

[48] M. Eitz, J. Hays, M. Alexa, "How do humans sketch objects?," ACM Trans. Graphics (TOG), 31(4): 1-10, 2012.

[49] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection," in Proc. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 1: 886-893, 2005.

[50] J. M. Saavedra, "Sketch based image retrieval using a soft computation of the histogram of edge local orientations (shelo)," in Proc. 2014 IEEE International Conference on Image Processing (ICIP): 2998-3002, 2014.

[51] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," Int. J. Comput. Vision, 60: 91-110, 2004.

Biographies

Azita Gheitasi received a Bachelor's degree in Electrical Engineering from Payam Noor University in Birjand center in 2014 and a Master's degree in Electrical Engineering from Shaukat Abad University in Birjand in 2017. Currently, she is a doctoral student in the field of Electrical Engineering and telecommunication Systems. Her research interests include image processing, deep learning, convolutional networks, cloud computing and machine learning.

- Email: A.gheitasi@birjand.ac.ir
- ORCID: 0009-0005-3953-9630
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Hassan Farsi received the B.Sc. and M.Sc. degrees from Sharif University of Technology, Tehran, Iran, in 1992 and 1995, respectively. Since 2000, he started his Ph.D. in the Centre of Communications Systems Research (CCSR), University of Surrey, Guildford, UK, and received the Ph.D. degree in 2004. He is interested in speech, image and video processing on wireless communications. Now, he works as Associate Professor in Communication Engineering in department of Electrical and Computer Eng., University of Birjand, Birjand, IRAN.

- Email: hfarsi@birjand.ac.ir
- ORCID: 0000-0001-6038-9757
- Web of Science Researcher ID: NA
- Scopus Author ID: 16202385600
- Homepage: https://cv.birjand.ac.ir/hasanfarsi/en

Sajad Mohamadzadeh received the B.Sc. degree in Communication Engineering from Sistan & Baloochestan, University of Zahedan, Iran, in 2010. He received the M.Sc. and Ph.D. degree in Communication Engineering from South of Khorasan, University of Birjand, Birjand, Iran, in 2012 and 2016, respectively. Now, he works as Associte Professor at department of Electrical and Computer Engineering, University of Birjand, Birjand, Iran. His area research interests include Image and Video Processing, Deep Neural Network, Pattern recognition, Digital Signal Processing, Sparse Representation, and Deep Learning.

- Email: s.mohamadzadeh@birjand.ac.ir
- ORCID: 0000-0002-9096-8626
- Web of Science Researcher ID: NA
- Scopus Author ID: 57056477500
- Homepage: https://cv.birjand.ac.ir/mohamadzadeh/en

How to cite this paper:
A. Gheitasi, H. Farsi, S. Mohamadzadeh, "Hybrid convolutional neural network with domain adaptation for sketch based image retrieval," J. Electr. Comput. Eng. Innovations, 12(2): 497-510, 2024.

DOI: 10.22061/jecei.2024.10778.735

URL: https://jecei.sru.ac.ir/article_2139.html