Deep CNN-based Features for Hand-Drawn Sketch Recognition via Transfer Learning Approach

Shaukat Hayat[1], Kun She*[2], Yao Yu[4]
School of Information and Software Engineering
University of Electronic Science and Technology of China
Chengdu, China

Muhammad Mateen[3]
School of Big Data and Software Engineering
Chongqing University, Chongqing
P.R China

Abstract—**Image-based object recognition is a well-studied topic in the field of computer vision. Features extraction for hand-drawn sketch recognition and retrieval become increasingly popular among the computer vision researchers. Increasing use of touchscreens and portable devices raised the challenge for computer vision community to access the sketches more efficiently and effectively. In this article, a novel deep convolutional neural network-based (DCNN) framework for hand-drawn sketch recognition, which is composed of three well-known pre-trained DCNN architectures in the context of transfer learning with global average pooling (GAP) strategy is proposed. First, an augmented-variants of natural images was generated and sum-up with TU-Berlin sketch images to all its corresponding 250 sketch object categories. Second, the features maps were extracted by three asymmetry DCNN architectures namely, Visual Geometric Group Network (VGGNet), Residual Networks (ResNet) and Inception-v3 from input images. Finally, the distinct features maps were concatenated and the features reductions were carried out under GAP layer. The resulting feature vector was fed into the softmax classifier for sketch classification results. The performance of proposed framework is comprehensively evaluated on augmented-variants TU-Berlin sketch dataset for sketch classification and retrieval task. Experimental outcomes reveal that the proposed framework brings substantial improvements over the state-of-the-art methods for sketch classification and retrieval.**

Keywords—Deep convolutional neural network; sketch recognition; transfer learning; global average pooling

I. Introduction

In a human point of view, sketch analysis is not only considering a fundamental problem, but it has a prominent role in the field of human-computer interaction (HCI). Sketches can be seen everywhere and have a significant role in daily life activities, i.e., education sector, art, design, and entertainment, etc. All through human society progress, the sketch has been utilized as a fundamental tool for conveying feelings, thoughts, judgments, and opinions. Since the ancient time, the people express their views in the form of sketch related petroglyphs or cave paintings. Such kind of art examples can be easily seen today in pre-historic art caves throughout the world.

The Technological explosion makes the sketches easy and ubiquitous. There exist for several fascinating multimedia applications such as HCI [1] and some other relevant work [2-4]. With the fame of touchscreens and smart-phone devices encouraged the people to draw sketches digitally. Presently, an excessive utilization of advanced technological tools, the one needs to access query sketch more accurately and retrieve its relevant contents to be well-recognized through technological-based smart devices. However, to acknowledge the needs of the society and to balance with technological advancement, the researchers have been analyzed various novel tasks regarding sketch recognition [5, 6], and sketch-based image retrieval [7, 8], in a field of computer vision. The idea behind the sketch classification or recognition is to extract the information from the desired object class of labeled sketch-images among the pre-defined set of object-classes. Based on the extracted information, the label of the targeted instance can be correctly identified. Classification or recognition techniques usually rely on extracted features through instance training before making a recognition. For sketch recognition, the researchers borrowed handcrafted features approaches which have been successfully used to extract features from natural images. There exist, Scale-Invariant Features Transforms (SIFT) [9] Histogram of Oriented Gradients (HOG) [10], descriptors and the bag-of-features has been already utilized. In this regards, different handcrafted features techniques are followed to yield the global features for sketch recognition, i.e., GF-HOG [11], FV [12], SSIM [13] and Structure Tensor [14]. Usually, handcrafted feature representations are not considered robust and also due to their high dimensionality make them computationally expensive. Current approaches to object recognition make the necessary use of deep learning and machine learning techniques. However, the most existing work in sketch recognition is based on deep learning approaches using deep convolutional neural networks (DCNNs) and showed an impressive result than handcrafted approaches [6, 8, 15].

In the recent past, deep learning frameworks based on DCNNs shows a breakthrough in different areas of computer vision, including vision recognition on large-scale challenging dataset [16, 17]. Moreover, deep learning approaches also benefitting sketch-based recognition and can provide useful features representations by analyzing large-scale sketch dataset, such as TU-Berlin sketch benchmark [18, 19]. Deep learning is capable of generating more distinctive features from sketch images and can leverage the performance for sketch classification or recognition as compared to use hand-crafted features. Deep features for sketch recognition was first time utilized by [20] and design a specialized neural network model. As a result, the classification accuracy on sketch image dataset TU-Berlin [21] has been improved as compare to hand-crafted features. Similarly, two different well-known CNNs models,

*Corresponding Author.

namely LeNet [22] and AlexNet [16] are used to extract features from sketch images and show improvement in the recognition results [15]. On the other hand, some recent attempts utilized different layers of various CNNs architectures for features extraction for the purpose of sketch classification and retrieval [8, 18, 23].

Visual recognition or classifications by deep learning approaches are mostly rely on extracted features. Generally, Deep CNNs features are categorized into three basic levels, such as high, middle, and low-level features. Each level of extracted features are having their strengths and potential in producing results and accuracy [24-26]. In order to obtain a higher recognition or classification accuracy and reducing the computational efforts, the concept of transfer learning (TL) approach can be exploited to get more robust features by combining the learned knowledge from multiple DCNNs models [27, 28]. In TL approach, first, the Deep CNNs models are trained on the generic visual dataset, and then pre-trained models can be directly used to train on domain-specific datasets. The motivation behind the TL approach is to combine more comprehensive and relevant knowledge of input objects resulting from multiple CNN architectures and then pass them through a classifier for a final decision. We believe that by doing so, it can achieve more robust and higher recognition accuracy as compared to the one extracted through single deep CNN model.

The sketches are mostly handled through smart-phones and other portable devices for different purposes in daily life routines. In this regard, we attempt to facilitate such touch-screen environment to retrieve the query sketch contents with higher recognition rate. To overcome the existing deficiencies in the sketch recognition system and following the emerging trend of exploring deep learning for features extraction via transfer learning approach, we proposed three different well-known robust DCNNs architectures in the state-of-the-art visual recognition to the task for sketch recognition. The proposed DCNNs architectures includes Inception-v3 [29] , ResNet [30] and VGGNet [17]. All these architectures have achieved promising performance on various challenges. These networks are trained on large-scale image dataset ImageNet [31].

The main contributions in this manuscript can be presented in the following:

- A novel and efficient CNN-based framework for hand-drawn sketch recognition is proposed that exploits the strength of extracted features from the various pre-trained DCNNs via transfer learning with the utilization of global average pool (GAP) concept.
- An attempt to generate the augmented-variants of natural images paired with TU-Berlin sketch dataset for enhancing a sketch recognition performance.
- A performance analysis of three individual deep CNNs architectures compare with proposed framework in a context of transfer learning with GAP for sketch recognition. The proposed framework obtained state-of-the-art recognition accuracy on augmented-variants TU-Berlin sketch dataset and also assesses on TU-Berlin sketch dataset (without augmented-variants).
- An evaluation of the proposed framework for sketch retrieval task.

The rest of the manuscript is organized as follows: In Section 2, we briefly present related literature based on handcrafted features and deep features. Section 3 describes the overall details of the proposed approach including data preparation and augmentation variants used in this study, the concept of transfer learning, and different proposed pre-trained deep CNNs architectures utilized in the current research. Section 4 provides results, analysis and evaluations of the proposed methodology. We conclude the manuscript in Section 5 along with the future directions.

II. Related Work

We include a review work for sketch recognition utilizing handcrafted-feature methods. Further, we enclosed our review details about deep learning approaches which have been used for hand-drawn sketch recognition and retrieval task. To hold focus, we threw light on the review work entirely related to hand-drawn sketch recognition.

A. Handcrafted Features

Previous sketch recognition problem was handled about CAD and artistic drawings by [32-34]. After releasing a large-crowed source TU-Berlin hand-drawn sketch dataset in 2012 by [21]. This dataset gains popularity among the computer vision researchers to utilize it further for recognition tasks. Variety of traditional approaches was carried out to classify different categories of sketch dataset and was tried to achieve higher recognition accuracy. Some researchers employed hand-engineered features techniques to extract the features for sketch recognition such as scale-invariant features transforms (SIFT) [9], histogram of oriented gradients (HOG) [10] and the bag-of-features techniques [35, 36].

Although, a method proposed by [21], describe the inter-class similarities and intra-class variations in large crowed-source sketch dataset. Support Vector Machine (SVM) classifier was used to learn the sketch representation in various object categories. Original sketch benchmark proposed by [21] and was then modified by [12]. The modified work uses SIFT, Gaussian Mixture Model (GMM) based fisher vector encoding for sketch recognition and fed into SVM classifier. This approach enhances the recognition performance near to human (73.1%) [21] accuracy rate against the same sketch dataset. A star graph based ensemble matching strategy was employed by [37], it covers not only local feature, but global structures of sketches were also adopted to match them. Further, structure matching was encapsulated, and bag-of-features was learned to exploit in a single framework. Eitz et al. [25] demonstrated hand-drawn sketch classification through implementing local features vectors techniques, i.e., SIFT and other different descriptors such as spark feature, shape context, HOG, and SHOG are embedded in a bag of features model and evaluate the performance on large scale sketch-based image dataset through Sketch-Based Image Recognition (SBIR) system. In [38], the author threw light on the proposed method Symmetric-aware Flip Invariant Sketch Histogram (SYM-

FISH) for sketch image retrieval and classification. Another approach of multi-kernel features was demonstrated by [39], where different local features were extracted to analyze the sketch image, integrate them to improve the sketch recognition performance. Individually, every feature performance was calculated and found that HOG outperformed as compare to others.

Different researchers have made the efforts through handcrafted features for sketch recognition, among these, a Fisher vector spatial pooling (SV-SP) [12], sketch image representation approach raises the sketch recognition performance up-to 68.9% to come close to 73.1% [21] human accuracy on TU-Berlin sketch benchmark. Generally, handcrafted features are not considered robust, and one of the limitation is high dimensionality of these features make them computationally expensive.

B. *Deep Features*

Recently, deep neural networks (DNNs) are utilized for various kind of problems, which have shown immense performance in different applications, including image recognition [16, 40, 41]. Deep networks have changed the trend by replacing hand-engineered features to the learning strategy. Instead of this, a wide range of research has been conducted comprising natural image recognition. AlexNet [16] outperforms on image recognition in comparative with others, and handle the ImageNet challenge with more significant improvements. Moreover, the utilization of deep neural networks has been expended to other tasks with variant sizes of network structures and depth according to the nature of the problem.

The networks, VGGNet [17], and GoogleNet [42] with deeper structure and ability to handle the complexity limitations of neural networks were introduced. The emergence of these deeper networks laid the foundation of a vast neural network named ResNet [30] having the residual connection, to permit the network for identity mapping tasks between the layers of the network. These deep neural networks were chosen and exercised on natural images to overcome the problem. However, several deep learning approaches have been adopted for sketch recognition. For the first time, an effort has been made to specially design a deep convolutional neural network (DCNN) architecture named sketch-DNN by [20]. Another research [15] extracts sketch features from two famous pre-trained CNNs, namely, AlexNet [16] and modified version of LeNet [22] and yield little improvement in the recognition results. The major contribution has been presented in [5], a deep CNNs model namely Skatch-a-Net was introduced for sketch recognition and beats the human sketch recognition accuracy. Later on, the existing model was modified in [6] and the sketch recognition performance gap increased from 1.8% to 4.9% than human recognition accuracy. Five convolutional layers CNN was trained by [43] by taking sketch images mixed with natural images as augmented training dataset. Further, to enhance the discriminative ability of the network, the training was presented with multiple rotated version of the sketch edge map and predicted the results with the labels. Jamil et al. [8] attempts to recognize partially colored hand-drawn sketch images and implemented fine-tuned CNN on augmented TU-Berlin sketch dataset to retrieve query-based sketch images through proposed model. The author [18] applied a feature fusion approach for sketch-based recognition system to considered different layers of CNNs for features extraction from the TU-Berlin sketch dataset.

III. Proposed Method

This section describes the proposed framework based on DCNNs architectures for sketch recognition. In the proposed method, three well-known Deep CNN architectures, i.e., Inception-v3 [29], ResNet [30], and VGGNet [17] in the state of the art of visual recognition for sketch analysis are used. The weights of these architectures are available for modification. These pre-trained models downloaded from the webpage of keras [44]. The weights are loaded to all the corresponding architectures. The proposed architectures are trained on augmented-variants TU-Berlin sketch dataset. The block diagram of the proposed framework is presented in Fig. 1.

A. *Data Preparation and Augmentation-Variants*

To carry out this experiment, a hand-drawn sketch dataset TU-Berlin [21] is utilized. The learning performance of deep convolutional neural networks depends on the availability of a large amount of training data. Data transformation and deformation techniques are used for expending the training dataset as an additional data samples to the existing labeled one, to reduce the overfitting problem. An essential concept of the data augmentation is that; the labels of the instances remain unchanged after applying this operation. Data augmentation can improve the generalization and discriminative ability of the model [16].

The most advanced augmentation method is adopted by mixing natural images with different transformations of enhanced edge, edge maps corresponding to the sketch images in the training dataset through anisotropic diffusion approach [45]. Fig. 2 illustrates natural image, edge enhanced and edge maps of natural image transformations. This will enable the proposed framework to compare effectively natural images; its various transformations i.e., edge maps and sketch images to match for the sketch recognition task. Different transformations and mixing natural images have been used by [8, 43] and extracted the features from both type of images i.e., sketch, natural images for recognition and retrieval task. In our case, the addition of augmentation-variants to the corresponding sketch objects categories will enable the network to learn more discriminative features representations. It will also facilitate the end-user to query sketch image through sketch retrieval system.

It is stated that augmented variants with sketch images will enhance the generalization ability of trained CNN-based framework on unseen sketch images.

Most likely, the edge-map exists in the hand-drawn sketch objects. To make it easy for the CNN-based framework to handle the edge maps of the natural images, the enhanced edges of natural images are formed and model them with edge maps of natural images. Gaussian smoothing method is utilized on the edge maps of natural images to form the enhanced-edge images of natural images. Mathematically, it can be presented as:

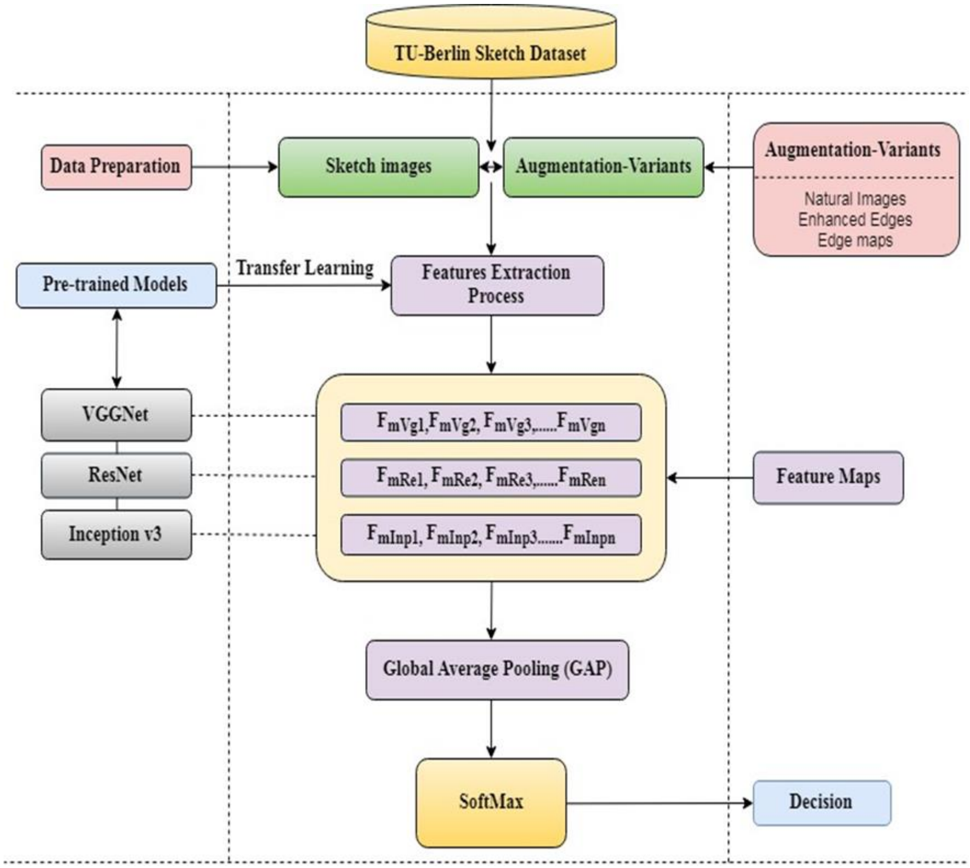

Fig. 1. Proposed Framework for Sketch Recognition.

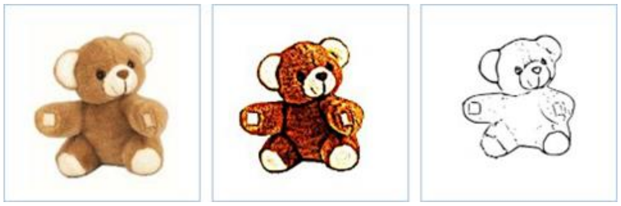

Fig. 2. Augmented-Variants of Natural Image for Training Proposed Framework, from Left to Right, the First Image is the Original (Natural Image), Second Sample Image is Enhanced Edge of Natural Image and the Last Represent the Edge Map of Natural Image.

$$I_{EE} = I(x,y) + \left[I(x,y) - \left[\frac{1}{2\pi\sigma^2} e^{\frac{-u^2+v^2}{2\sigma^2}} * I(x,y) \right] \right] \tag{1}$$

where, I represent input image, IEE indicate enhanced edge image, and x and y are the special coordinates for the image, σ = 0.5 represent standard deviation for filter and * is the convolution operation. The geometric transformation techniques are applied, i.e., flips and rotations on augmented-variants to rich the training samples and to avoid the overfitting problem.

To this end, the symbolic notations are assigned to represent the training data of sketch images, natural images, and other augmented-variants, i.e., edge maps and enhanced edges for the proposed framework. The sketch images are represented as:

$$S_{img_n} = \left\{ S_{img_1}, S_{img_2}, ...S_{img_n} \right\} \in \square^{1\times n} \tag{2}$$

where, n denoted the number of training sketch images, similarly, for natural images;

$$N_{img_m} = \left\{ N_{img_1}, N_{img_2}, ...N_{img_m} \right\} \in \square^{1\times m} \tag{3}$$

here, m shows the number of natural images for training. The labels assign to the natural images;

$$L = \left\{ l_1, l_2, ...l_n \right\} \in \square^{c\times m} \tag{4}$$

and c shows the categories. Furthermore, the edge-maps and enhanced edge images generated from natural images and added to the relevant categories of natural images. Finally, natural images with augmented variants are sum-up to the corresponding sketch images within specified sketch object categories, i.e.

$$S_{img_n} N_{img_t} = \left\{ S_{img_1} .N_{img_1}, S_{img_2} .N_{img_2}, ...S_{img_n} .N_{img_t} \right\} \in \square^{n\times t\times c} \tag{5}$$

where, n and t are the sketch and natural images with augmented-variants respectively and c represents the corresponding object category, for training the proposed framework.

B. *Pre-Trained CNNs Architectures*

Several state-of-the-art deep neural networks (DNNs) have been utilized for various kind of problems and gives outstanding performance in the field of computer vision application such as classification, recognition, etc [46, 47]. In the proposed methodology for sketch recognition, three different pre-trained deep CNNs models are adopted for features extraction via transfer learning, includes Inception-v3 [29], ResNet [30] and VGGNet [17]. Although the architectures of these networks are different from one another, each of the adapted model architecture is describe in the following:

1) Inception-v3: Inception-v3[29] is a deep convolutional neural network and the winner of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC)-2014. This architecture has been implemented as an updated version of the GoogleNet [42]. with a depth of 44-layers and 21 million learnable parameters. Inception module is illustrated in Fig. 3.

2) ResNet: ResNet CNN is a very deep residual network, proposed by He et al.[30] to addresses the training problems confronted by deep CNNs. This model received promising results on ImageNet. The complexity of this network is higher than other CNNs architectures due to the existence of its 152 layers. Shortcut connections are one of the critical innovation of ResNet CNN, which carries out the identity mapping, and their output is linked to stacked layers' output. ResNet CNN is illustrated in Fig. 4.

3) VGGNet: VGGNet is a CNN model, invented by visual geometry group (VGG) of Oxford University [17]. This model is the first runner-up of ILSVR-2014 for classification and the winner of localization task. The architecture of VGGNet is similar to AlexNet, and the only difference is the depth of the VGGNet. This architecture consists of 19 layers, including convolution, pooling, and three fully-connected layers. The network consists of small convolutional kernel 3x3 with stride 1. It performs better than AlexNet. The architecture of the VGGNet is shown in Fig. 5.

C. *Transfer Learning*

In the context of traditional machine learning algorithms, it is assumed that the characteristics of features space based on training and testing data are equal [48]. However, in a practical world, such a big amount of data is not cheap and also very hard to collect. Transfer learning is the reasonable solution to tackle such kind of problems and can provide an accurate result with less training samples.

Transfer Learning technique is widely used in the machine learning to utilize useful information from the set of source point to the set of target point [49]. The inspiration behind the adaptation of transfer learning is to solve a problem with improved results for the target domain. To be more specific, for example, the base model is first trained on relevant data instances for a specific task and then move to the target task trained by their data instances [48]. Transfer learning is the best choice for the case when the dataset of the source domain is bigger than the dataset of the target domain. If the size of the dataset for the target domain is smaller and similar to the dataset of source domain, then overfitting possibility is high. Alternatively, the chance of overfitting is reduced, and only fine-tuning of the pre-trained model is required if the size of data for the target domain is large and similar to the dataset of the source domain.

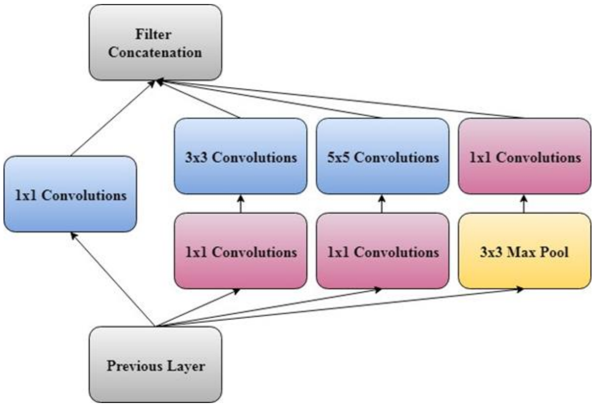

Fig. 3. Basic Inception Module.

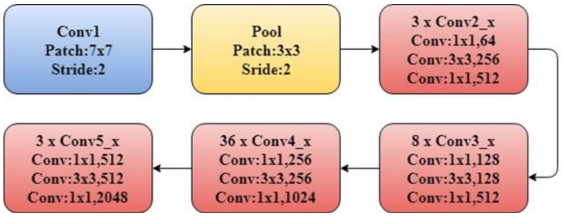

Fig. 4. Basic Architecture of ResNet.

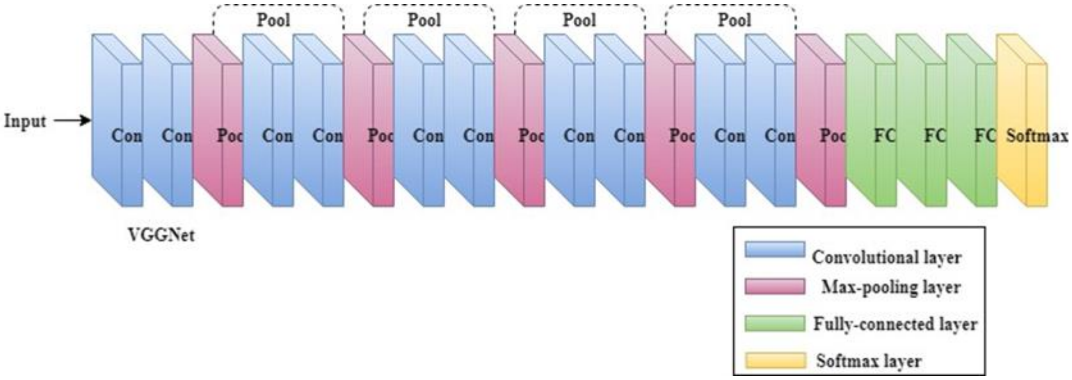

Fig. 5. The basic Architecture of VGGNet.

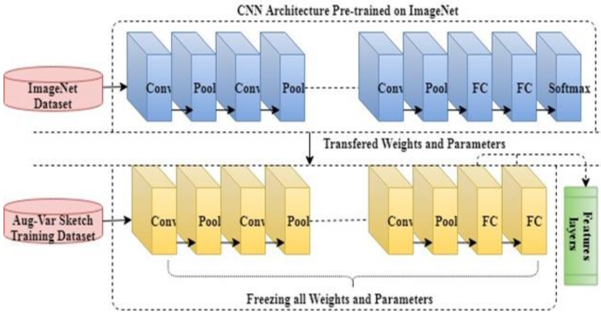
CNN Architecture Pre-trained on ImageNet
ImageNet Dataset
Conv
Pool
Conv
Pool
Conv
Pool
FC
FC
Softmax
Transfered Weights and Parameters
Aug-Var Sketch Training Dataset
Conv
Pool
Conv
Pool
Conv
Pool
FC
FC
Features layers
Freezing all Weights and Parameters

Fig. 6. Transfer Learning Concept.

In this study, three different deep CNNs models are utilized and pre-trained on ImageNet (Source domain) [50]. Based on trained knowledge, these models re-used for sketch dataset (Target domain) and fine-tuned. The concept of transfer learning is shown in Fig. 6. After that, the transfer learning approach is adopted. By doing so, it enables the models to learn more generic features from other datasets without the efforts to conduct new training. So, these architectures are trained on augmented-variants sketch dataset. The resulting feature maps from the proposed deep CNNs architectures are then concatenated into a global average pooling (GAP) layer to generate a feature vector for sketch classification.

D. Global Average Pooling

For sketch classification, proposed deep CNNs architectures extract the high-level features maps by taking advantage of convolutional layers as a features extractor. Since, as compared to single model utilization for feature maps generation, a different feature maps from the different architectures demonstrate diverse characteristics for input patterns. In this case, these distinct high-level feature maps are concatenated to maintain the discriminative knowledge about the input data. Concatenation of feature maps from different architectures can create a curse-of-dimensionality. To overcome this problem, a global average pooling (GAP) layer [51] is applied to replace all of the fully connected layers in proposed deep CNNs architectures on top of the feature maps. All the extracted feature maps are concatenated into GAP layer. This layer takes the average of each feature map and generates the features vector as the output of the GAP layer to directly fed into the softmax classifier for each corresponding sketch object category. One of the advantages of this layer is the summarization of spatial information. The resulting features vector of GAP is more robust to a spatial translation of the input images. It can reduce the total number of parameters in the network and perform dimensionality reduction to enhance the generalization ability of the networks. However, the overfitting problem can be reduced automatically without optimizing any parameter in the GAP layer.

IV. Results and Analysis

In this section, the proposed framework is evaluated for sketch recognition. The detailed description of the dataset is provided for the utilization and validation of the proposed method for sketch recognition problem. Further, the achieved results are provided and compared with state-of-the-art methods.

A. Dataset

To evaluate the performance of the proposed method for sketch recognition, a TU-Berlin sketch dataset [21] is used. This dataset consists of total 20,000 sketch images distributed over 250 object categories. Each object category is having 80 sketch objects. Total 1,350 non-experts sketch drawers take participation in the sketch generation event conducted by Amazon Mechanical Turk (AMT) and was generated with aim of hand-drawn sketch recognition and classification purposes. The human recognition performance on this dataset is 73.1%. The size of each sketch image is 1111x1111. Few of sketch image samples from different object classes are shown in Fig. 7.

B. Natural Images for Augmented-Variants

To get the color images for augmented variants, the natural photos are collected from the publically available full-colored image dataset and most of them collected from the web. These dataset covers some of the object categories of TU-Berlin dataset. These images were collected from Caltech 256 image dataset [52], while the remaining images were taken from the web and generates the augmented variants, i.e., enhanced edges and edge maps of natural images. Finally, Total 31,500 colored images, including other augmented-variants, are collected. These images are then added to the corresponding sketch object categories of TU-Berlin sketch dataset for training the proposed framework. Some of the sample images corresponding to sketch images are shown in Fig. 8. The evaluations are carried out with 3-fold cross validation to allow comparisons with baselines.

C. Environment Setting

The proposed framework is implemented in open source keras using python libraries. It consists of three different pre-trained deep CNN architectures which are trained separately to extract the features from the augmented-variants sketch dataset. The overall training is conducted on NVIDIA dual Xeon processor with 13GB RAM and GPU cards. Ubuntu 16.04 operating system with the 64bit environment is used to perform training operations.

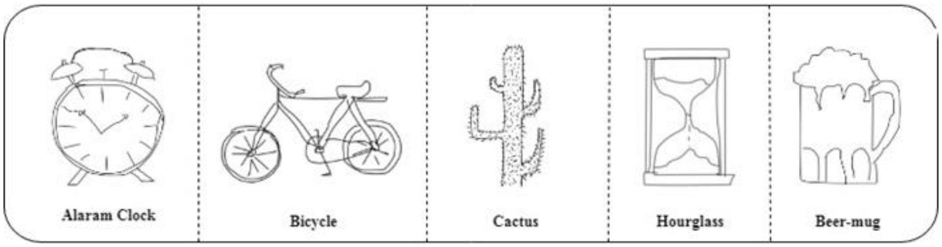

Fig. 7. TU-Berlin Sample Sketch Images.

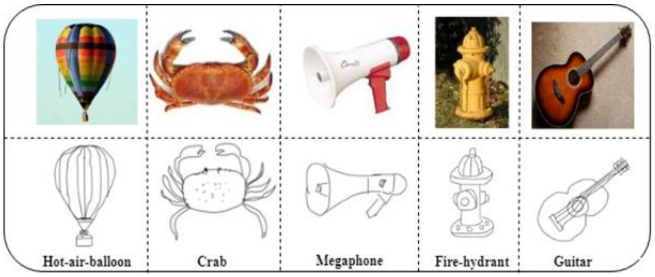

Fig. 8. Sample Colored Natural Images with Corresponding Sketch Image.

D. Results and Evaluation

In order to validate the performance of the proposed CNN-based framework for sketch recognition, the sketch classification accuracies are shown in Table I. According to tabulated outcomes, the performance of proposed individual CNN architectures, i.e., VGGNet, ResNet, and Inception-v3 obtained 78.93%, 89.61%, and 91.89% classification accuracy, respectively. However, the proposed framework achieves better performance with 94.57% sketch classification accuracy on augmented-variants TU-Berlin sketch dataset and beat the individuals, i.e., VGGNet, ResNet and Inception-v3 in performance gap with 15.64%, 4.96%, and 2.68%, respectively. It is evident from the tabulated results that proposed method outperforms in terms of sketch classification than the performance of other three individual architectures.

In this subsection, the performance of proposed framework is compared with state-of-the-art methods including sketch-based handcrafted features and deep features methods. The overall accuracy is shown in terms of percentage to demonstrate the results. Fig. 9 shows the hand-crafted features recognition performances on sketch images. The proposed method outperforms on HOG-SVM (recognition accuracy 56.0%), sketch recognition accuracy achieved through Ensemble method, Multi-kernel-SVM and Fisher Vector-SP were 61.5%, 65.8%, and 68.9%, respectively in the literature study.

HOG based features with SVM classifier has the lowest recognition rate. However, the best performance accuracy based on hand-crafted features is 68.9%, which is less than the human recognition accuracy (73.1%) on sketch data. The reason of lower performance of handcrafted features is that; mostly these methods have been designed to extract features from real photos and not suitable to covers the high variability of abstractions and appearance in sketch images.

Similarly, the results for deep features methods are summarized in Table II. It shows that deep networks perform better than hand-crafted features. Human recognition level accuracy on TU-Berlin dataset was first beats through deep network architecture [5] and enhance recognition rate with 1.8% higher than human recognition performance. Moreover, the performance gap grows from 1.8% to 4.2% when [43] implemented deep sketch model by mixing sketch images with colored images for sketch recognition. The sketch recognition accuracy 79.1% achieved by [8] using pre-trained VGGNet architecture through transfer learning approach.

Different reasons might cause variations in sketch recognition results while using DNNs model. It depends on the structure of the model, the depth of network architecture, different methods used for feature extractions, and tuning-up various parameters. Even though, to check the effect of every parameter for any model performance is also arduous and tedious task.

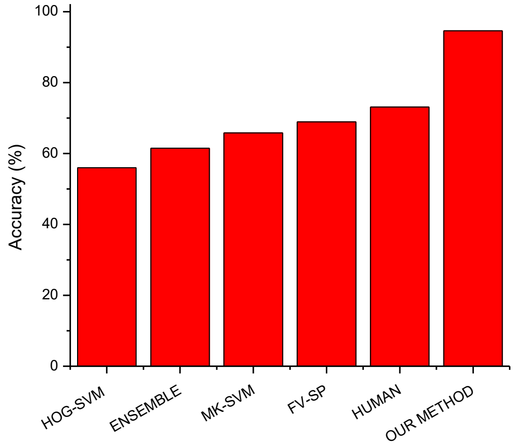

Fig. 9. Sketch Classification Accuracy based-on Hand-Crafted Features.

TABLE. I. COMPARATIVE SKETCH CLASSIFICATION ACCURACY OF THE PROPOSED FRAMEWORK WITH OTHER INDIVIDUAL DCNNS ARCHITECTURES

Index	Method	Accuracy
1	VGGNet	78.93%
2	ResNet	89.61%
3	Inception-v3	91.89%
4	**Proposed Method with GAP**	**94.57%**

Therefore, some conclusions can be made from the baseline results mentioned in the Fig. 9 and Table II. First of all, the proposed CNNs-based framework consistently outperforms in sketch classification on both handcrafted features methods and the sketch features analyzed through deep neural network models. This can show that the use of natural images plays a significant role in the evaluation of sketch images. Additionally, various augmentation-variants, specifically edge maps strengthen the proposed framework capability in sketch recognition accuracy. Secondly, the performance of individual deep CNNs becomes improved when it goes deeper. The best recognition performance through individual deep CNN architecture is achieved by Inception-v3. Thirdly, in a case of transfer learning approach, the distinct features of three deep CNNs architectures were combined, and the recognition performance is improved by employing GAP strategy. This outcome substantiates the experiments for augmented-variants TU-Berlin sketch dataset. In the proposed framework, it is declared that using transfer learning with GAP increases sketch recognition accuracy. This is also applicable for combining distinctly extracted features from multiple CNNs architectures as compared to individual CNN architecture.

TABLE. II. SKETCH CLASSIFICATION ACCURACY COMPARISON BASED ON DEEP FEATURES

Index	Methods	Accuracy
1	AlexNet-SVM [16]	67.1%
2	AlexNet-Sketch [16]	68.6%
3	LeNet [22]	55.2%
4	Human [21]	73.1%
5	Sketch-a-Net [5]	74.9%
6	Deepsketch [43]	77.3%
7	VGG-based Transfer Learning [8]	79.1%
8	**Proposed Method with GAP**	**94.57%**

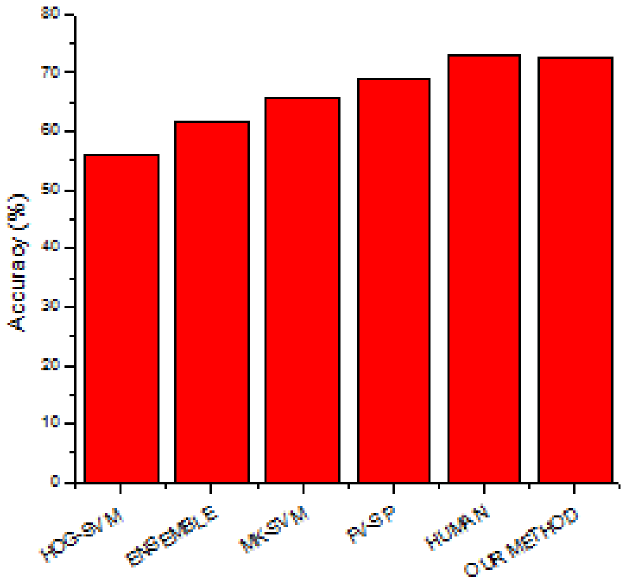

Fig. 10. Sketch Classification Results Comparison on TU-Berlin (without Augmented Variants).

On the other hand, we evaluate our proposed method on TU-Berlin sketch dataset (without augmented-variants), the experimental results are illustrated in Fig. 10. The proposed framework achieves a competitive performance of 72.82% classification accuracy as compared with 73.1% human recognition accuracy except those sketch based CNNs architectures [5, 7], which have been specifically designed for sketch classification.

E. Further Evaluation for Retrieval Task

The performance of the proposed deep framework is further evaluated on sketch retrieval task. For this test, the proposed deep CNNs-based frame work is used to extract features from both the sketch images and natural images. All the images are indexed with concerned features. For retrieval task, proposed framework is used to extract the features from the edge maps and query sketches separately and compared with all the retrieval candidate images in the database. Euclidean distance is computed to make the comparison between the query sketch images and the images in the database. The query images are randomly selected to retrieve the similar images from the image database.

The sketch object-based queries and top-9 retrieval results are shown in Fig. 11. The retrieval results are ranked with scored values. The lower scored value represents the higher rank similarity between the query sketch and the retrieved image. In most of the cases, the query images retrieved the most similar candidate images which show the enough discriminative features for retrieval task.

These images (i.e., teapot, beer-mug, guitar, etc.) are retrieved with high rank similarity. Interestingly, the retrieved image had very comparable edge maps which make the retrieval task in high ranks. Moreover, in some cases the system failed to retrieve the right candidate images, the reason might be the natural images having complex background leading the large difference between the sketch images and edge-maps.

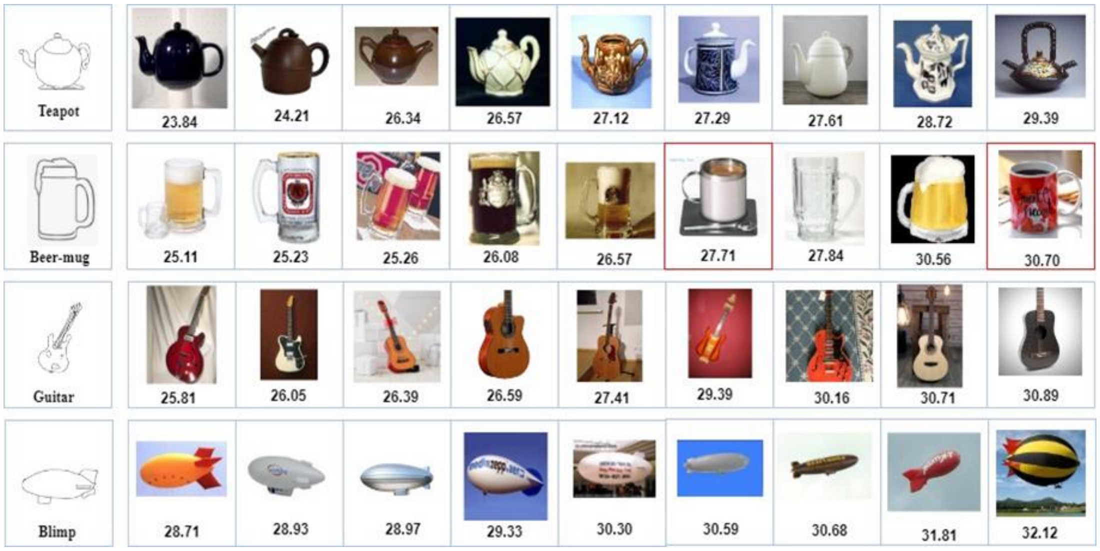

Fig. 11. Retrieval Performance of our SBIR, Top-9 Retrieval Outcomes of Four Sketch Queries by Proposed CNNs-based Framework. The First Column Indicates the Query Sketches, and the Most Similar Retrieved Candidate Images According to Ranked-Score from Four Different Object Categories are Shown From the Top to the Bottom-Row in the Sequence. Red-Box Images Indicate Incorrect Retrieved Images.

The capabilities of the proposed framework are extensively evaluated on a retrieval task. The lower rank similarity retrieval results with the corresponding scored values are illustrated in Fig. 12. Additionally, the retrieval performance of the proposed framework was also tested on sketches and natural images which was not a part of training or validation data. It is interesting to demonstrate that the query sketches retrieve the nearest candidate images with less ambiguity. However, the performance can be further improved by providing enough training data instances and also by reducing background complexity of the natural images. Results illustrated in Fig. 13 advocate that proposed CNN-based framework is capable to perform well on the variety of images which was not a part of training or validation data. These experiments validate the effectiveness of the proposed framework for sketch recognition.

F. *Experimental Analysis*

The proposed CNNs-based framework outperforms all baselines and achieves better performance on TU-Berlin (augmented-variants) sketch dataset which shows that augmented-variants are beneficial for sketch recognition. It is worth mentioning that handcrafted features performance on this dataset is worse than deep feature methods. In our case, the complex background of the natural images and generated edge maps from those images, make the proposed method more challenging and competitive. The proposed method performed better and retrieved the candidate images mostly in high ranked score of similarity. This demonstrate that proposed framework based on transfer learning with GAP is capable to extract the most discriminative features from both type of images i.e., sketch and natural images and could help to strengthen retrieval performance. But Fig. 12 and Fig. 13 represent the retrieval results of lower rank similarity and the results for images instead of using training and validation images respectively, where the incorrect retrieved images are outlined in red-boxes. Therefore, it is stated that the lower rank similarity performance and incorrect retrieved images might be the reasons of providing not enough training samples to the proposed framework as well as it might be not well-aligned with complex background of candidate retrieval images.

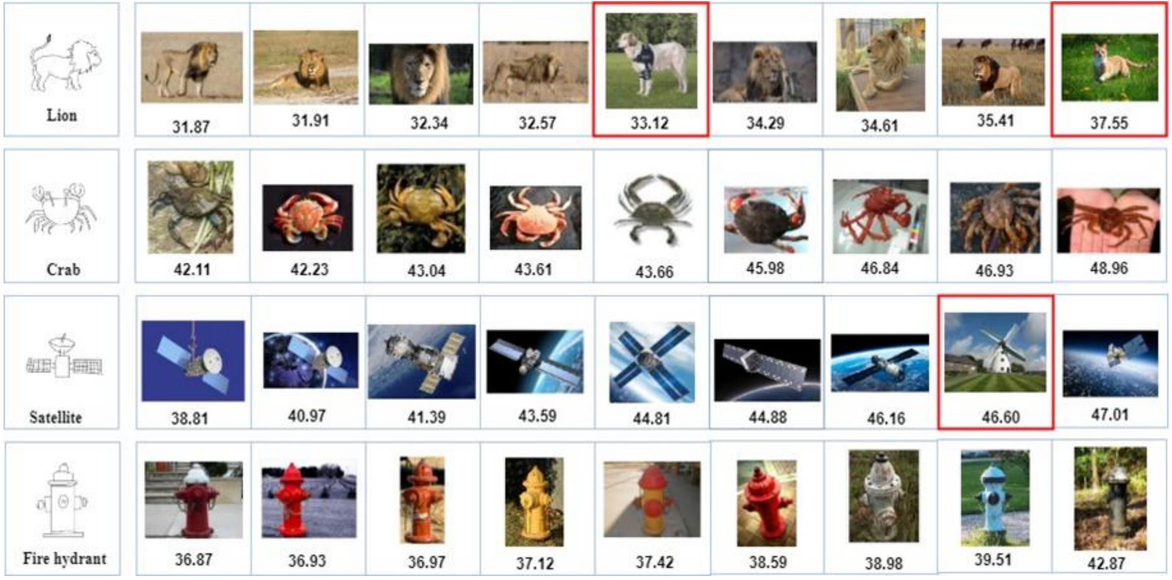

Fig. 12. Low-Rank Similarity Retrieval Performance of SBIR. The Red-Outlined Boxes Indicate Incorrect Retrieved Candidate Objects.

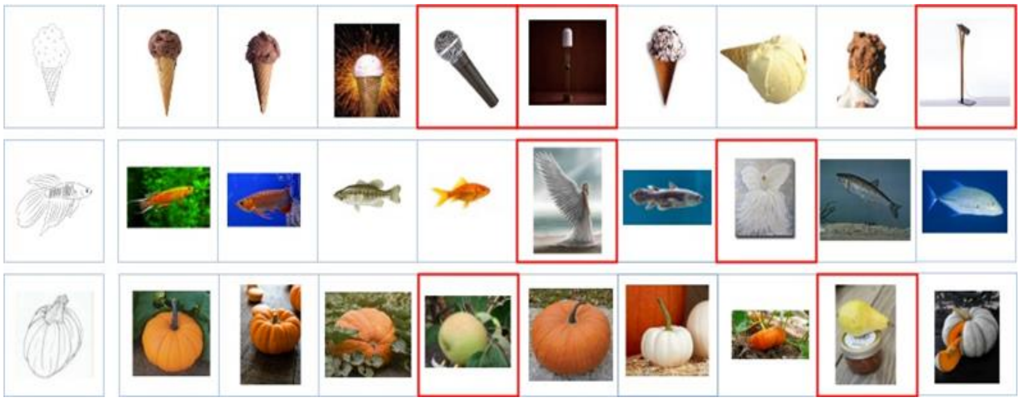

Fig. 13. Retrieval Performance based on Sketch-Query for other Object Categories.

V. Conclusions

In this manuscript, a deep CNN-based framework for sketch recognition via transfer learning with global average pooling strategy was proposed. The three different well-known pre-trained DCNNs models are analyzed in the proposed framework and fine-tuned on both with and without augmented-variants TU-Berlin sketch dataset. The sketch classification performance of the proposed framework was compared with different CNN architectures individually and also compared with other state-of-the-art methods. Considering the individual CNN architectures, it is observed from the experimental results based on augmented variants TU-Berlin sketch dataset that; the Inception-v3 showed higher accuracy than other two, i.e., ResNet and VGGNet architectures. However, VGGNet showed lowest classification accuracy among the other individual CNN architectures. The proposed framework outperformed the other existing methods in terms of sketch classification and retrieval task. The utilization of GAP not only reduces feature dimensionality but, also enhances the classification accuracy over the augmented-variants TU-Berlin sketch dataset. On the other hand, the proposed framework provides a competitive result as compare with human recognition accuracy on (without augmented-variants) TU-Berlin sketch dataset. In the future, deep learning approaches would be adopted for 3D shapes object retrieval task.

References

[1] D. Dixon, M. Prasad, and T. Hammond, “iCanDraw: using sketch recognition and corrective feedback to assist a user in drawing human faces,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Atlanta, Georgia, USA, 2010, pp. 897-906.

[2] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view Convolutional Neural Networks for 3D Shape Recognition." pp. 945-953.

[3] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan, "Deep Human Parsing with Active Template Regression," arXiv e-prints, https://ui.adsabs.harvard.edu/abs/2015arXiv150302391L, [March 01, 2015, 2015].

[4] M. Eitz, R. Richter, K. Hildebrand, T. Boubekeur, and M. Alexa, “Photosketcher: Interactive Sketch-Based Image Synthesis,” IEEE Computer Graphics and Applications, vol. 31, no. 6, pp. 56-66, 2011.

[5] Q. Yu, Y. Yang, Y.-Z. Song, T. Xiang, and T. Hospedales, "Sketch-a-Net that Beats Humans," arXiv e-prints, https://ui.adsabs.harvard.edu /abs/2015arXiv150107873Y, [January 01, 2015, 2015].

[6] Q. Yu, Y. Yang, F. Liu, Y.-Z. Song, T. Xiang, and T. M. Hospedales, “Sketch-a-Net: A Deep Neural Network that Beats Humans,” International Journal of Computer Vision, vol. 122, no. 3, pp. 411-425, May 01, 2017.

[7] O. Seddati, S. Dupont, and S. Mahmoudi, “DeepSketch 3,” Multimedia Tools and Applications, vol. 76, no. 21, pp. 22333-22359, November 01, 2017.

[8] J. Ahmad, K. Muhammad, and S. W. Baik, “Data augmentation-assisted deep learning of hand-drawn partially colored sketches for visual search,” PloS one, vol. 12, no. 8, pp. e0183838, 2017.

[9] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110, November 01, 2004.

[10] N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection." pp. 886-893 vol. 1.

[11] R. Hu, S. James, T. Wang, and J. Collomosse, “Markov random fields for sketch based video retrieval,” in Proceedings of the 3rd ACM conference on International conference on multimedia retrieval, Dallas, Texas, USA, 2013, pp. 279-286.

[12] R. G. Schneider, and T. Tuytelaars, “Sketch classification and classification-driven analysis using fisher vectors,” ACM Transactions on Graphics (TOG), vol. 33, no. 6, pp. 174, 2014.

[13] E. Shechtman, and M. Irani, "Matching Local Self-Similarities across Images and Videos." pp. 1-8.

[14] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, “A descriptor for large scale image retrieval based on sketched feature lines,” in Proceedings of the 6th Eurographics Symposium on Sketch-Based Interfaces and Modeling, New Orleans, Louisiana, 2009, pp. 29-36.

[15] R. Kiran Sarvadevabhatla, and R. Venkatesh Babu, "Freehand Sketch Recognition Using Deep Features," arXiv e-prints, https://ui.adsabs.harvard.edu/abs/2015arXiv150200254K, [February 01, 2015, 2015].

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks." pp. 1097-1105.

[17] K. Simonyan, and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv e-prints, https://ui.adsabs.harvard.edu/abs/2014arXiv1409.1556S, [September 01, 2014, 2014].

[18] E. Boyaci, and M. Sert, "Feature-level fusion of deep convolutional neural networks for sketch recognition on smartphones." pp. 466-467.

[19] Y. Qi, Y.-Z. Song, H. Zhang, and J. Liu, "Sketch-based image retrieval via siamese convolutional neural network." pp. 2460-2464.

[20] Y. Yang, and T. Hospedales, Deep Neural Networks for Sketch Recognition, 2015.

[21] M. Eitz, J. Hays, and M. Alexa, “How do humans sketch objects?,” ACM Trans. Graph., vol. 31, no. 4, pp. 1-10, 2012.

[22] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, 1998.

[23] J. Bai, M. Wang, and D. Kong, “Deep Common Semantic Space Embedding for Sketch-Based 3D Model Retrieval,” Entropy, vol. 21, no. 4, pp. 369, 2019.

[24] N. Upadhyaya, and M. Dixit, A Review: Relating Low Level Features to High Level Semantics in CBIR, 2016.

[25] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, “Sketch-based image retrieval: Benchmark and bag-of-features descriptors,” IEEE transactions on visualization and computer graphics, vol. 17, no. 11, pp. 1624-1636, 2010.

[26] J. Yang, S. Li, and W. Xu, “Active Learning for Visual Image Classification Method Based on Transfer Learning,” IEEE Access, vol. 6, pp. 187-198, 2018.

[27] J. T. Zhou, S. J. Pan, and I. W. Tsang, “A deep learning framework for Hybrid Heterogeneous Transfer Learning,” Artificial Intelligence, 2019/06/06/, 2019.

[28] L. Zhang, D. Wang, C. Bao, Y. Wang, and K. Xu, “Large-Scale Whale-Call Classification by Transfer Learning on Multi-Scale Waveforms and Time-Frequency Features,” Applied Sciences, vol. 9, no. 5, pp. 1020, 2019.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. B. Wojna, Rethinking the Inception Architecture for Computer Vision, 2016.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," arXiv e-prints, https://ui.adsabs.harvard.edu/abs/ 2015arXiv151203385H, [December 01, 2015, 2015].

[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database." pp. 248-255.

[32] M. Rahim, N. Othman, and Z. Jupri, “A Comparative Study on Extraction and Recognition Method of CAD Data from CAD Drawings, Information Management and Engineering,” ICIME, vol. 9, pp. 709-713.

[33] C. L. Zitnick, and D. Parikh, "Bringing semantics into focus using visual abstraction." pp. 3009-3016.

[34] P. Sousa, and M. J. Fonseca, “Geometric matching for clip-art drawing retrieval,” Journal of Visual Communication and Image Representation, vol. 20, no. 2, pp. 71-83, 2009.

[35] A. McCallum, and K. Nigam, "A comparison of event models for naive bayes text classification." pp. 41-48.

[36] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features." pp. 137-142.

[37] Y. Li, Y.-Z. Song, and S. Gong, "Sketch Recognition by Ensemble Matching of Structured Features." p. 2.

[38] X. Cao, H. Zhang, S. Liu, X. Guo, and L. Lin, "Sym-fish: A symmetry-aware flip invariant sketch histogram shape descriptor." pp. 313-320.

[39] Y. Li, T. M. Hospedales, Y.-Z. Song, and S. Gong, “Free-hand sketch recognition by multi-kernel feature learning,” Computer Vision and Image Understanding, vol. 137, pp. 1-11, 2015.

[40] N. Mboga, S. Georganos, T. Grippa, M. Lennert, S. Vanhuysse, and E. Wolff, “Fully Convolutional Networks and Geographic Object-Based Image Analysis for the Classification of VHR Imagery,” Remote Sensing, vol. 11, no. 5, pp. 597, 2019.

[41] D. Mao, and Z. Hao, “A Novel Sketch-Based Three-Dimensional Shape Retrieval Method Using Multi-View Convolutional Neural Network,” Symmetry, vol. 11, no. 5, pp. 703, 2019.

[42] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," arXiv e-prints, https://ui.adsabs.harvard.edu/abs/2014arXiv1409.4842S, [September 01, 2014, 2014].

[43] X. Wang, X. Duan, and X. Bai, “Deep sketch feature for cross-domain image retrieval,” Neurocomput., vol. 207, no. C, pp. 387-397, 2016.

[44] “https://keras.io/applications/.”

[45] P. Perona, and J. Malik, “Scale-space and edge detection using anisotropic diffusion,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 7, pp. 629-639, 1990.

[46] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, “Deep Pyramidal Residual Networks for Spectral–Spatial Hyperspectral Image Classification,” IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 2, pp. 740-754, 2019.

[47] J. Kim, B. Kim, P. P. Roy, and D. Jeong, “Efficient Facial Expression Recognition Algorithm Based on Hierarchical Deep Neural Network Structure,” IEEE Access, vol. 7, pp. 41273-41285, 2019.

[48] K. Weiss, T. M. Khoshgoftaar, and D. Wang, “A survey of transfer learning,” Journal of Big Data, vol. 3, no. 1, pp. 9, May 28, 2016.

[49] S. J. Pan, and Q. Yang, “A Survey on Transfer Learning,” IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, 2010.

[50] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” International Journal of Computer Vision, vol. 115, no. 3, pp. 211-252, December 01, 2015.

[51] M. Lin, Q. Chen, and S. Yan, "Network In Network," arXiv e-prints, https://ui.adsabs.harvard.edu/abs/2013arXiv1312.4400L, [December 01, 2013, 2013].

[52] “https://www.kaggle.com/jessicali9530/caltech256 ”.