Noname manuscript No.
(will be inserted by the editor)

Recent Advances in Scene Image Representation and Classification

Chiranjibi Sitaula* · Tej Bahadur Shahi · Faezeh Marzbanrad · Jagannath Aryal

Received: DD Month YEAR / Accepted: DD Month YEAR

Abstract With the rise of deep learning algorithms nowadays, scene image representation methods have achieved a significant performance boost, particularly in accuracy, in classification. However, the performance is still limited because the scene images are mostly complex having higher intra-class dissimilarity and inter-class similarity problems. To deal with such problems, there have been several methods proposed in the literature with their advantages and limitations. A detailed study of previous works is necessary to understand their advantages and disadvantages in image representation and classification problems. In this paper, we review the existing scene image representation methods that are being widely used for image classification. For this, we, first, devise the taxonomy using the seminal existing methods proposed in the literature to this date using deep learning (DL)-based, computer vision (CV)-based, and search engine (SE)-based methods. Next, we compare their performance both qualitatively (e.g., quality of outputs, pros/cons, etc.) and quantitatively (e.g., accuracy). Last, we speculate on the prominent research directions in scene image representation tasks using

Corresponding Author (*C. Sitaula) · F. Marzbanrad
Department of Electrical and Computer Systems Engineering
Monash University
Wellington Rd, Clayton VIC 3800, Australia
E-mail: chiranjibi.sitaula@monash.edu

TB Shahi
School of Engineering and Technology
Central Queensland University, Rockhampton, QLD, 4701, Australia
and
Central Department of Computer Science and Information Technology (CDCSIT)
Tribhuvan University
TU Rd, Kirtipur 44618, Kathmandu, Nepal

J. Aryal
Department of Infrastructure Engineering
The University of Melbourne
Parkville VIC 3010, Australia

keyword growth and timeline analysis. Overall, this survey provides in-depth insights and applications of recent scene image representation methods under three different methods.

Keywords Computer vision · Classification · Deep learning · Machine learning · Scene image representation

1 Introduction

Scene image analytics (e.g., scene representation, classification, clustering, etc.) is a highly-researched topic owing to its strong connection to recent technologies such as sensors, video cameras, robotics, and the internet of things (IoT) [1]. It also has an association with other sectors such as hyperspectral image analytics [2], satellite image analytics [3], climate image analytics [4], and so on. The image representation methods for each of them are dependent on the nature of the images; therefore, we need to adopt the appropriate feature extraction methods for their representation accordingly [5]. To perform such tasks, researchers have extended their works from very basic levels that use traditional computer vision-based methods to more sophisticated levels that use recent deep learning-based methods in addition to search engine-based methods.

Initially, researchers mostly preferred to use the traditional Com,puter Vision (CV)-based methods until 2014 for the scene image representation tasks. This is because Deep Learning (DL) models did not flourish at that time and traditional CV-based methods dominated scene representation tasks. Later on, DL-based methods, which originated in 1943 [6], have been dominant in the computer vision community from 2014 until now, particularly for scene image representation and classification [1]. Recently, to tackle the weaknesses of visual information achieved from either traditional CV-based methods or DL-based methods, in 2019, researchers proposed new methods based on the Search Engine (SE) to capture the contextual information for the scene image representation tasks, which are also called SE-based methods [7].

Because of such predominant growth and application of such methods, it has been challenging to explore the potential of each of them. Therefore, a survey study is crucial, not only to explore the surging potentials but also to help understand the application areas, research trends, and developments. Some recent review works

Questions	Wei et al. [8]	Anu et al. [9]	Singh et al. [10]	Xie et al. [11]	**Ours**
Traditional CV-based methods?	✓	✓	✓	✓	✓
Latest DL-based methods?	✗	✗	✗	✓	✓
SE-based methods?	✗	✗	✗	✗	✓
Trend and keyword growth analysis?	✗	✗	✗	✗	✓

Table 1: Comparison of our work with existing works

related to scene image representation are summarised below, whereas the summary is reported in Table 1.

(i) Wei et al. [8] studied the traditional feature extraction methods using empirical analysis, when the DL-based methods were not dominant, which helped understand the efficacy of traditional feature extraction methods for scene image representation. In addition, they perform an empirical study of such methods on four benchmark datasets. However, they explain limited DL-methods for scene image representation, which lacks in-depth elaboration of recent DL methods in this domain.
(ii) Anu et al. [9] discussed the traditional CV-based methods to extract the image features, which shed light on the applicability of different CV-based methods for scene image representation during that time. However, their study does not classify the traditional CV-based methods in-detail.
(iii) Singh et al. [10] presented a review of recent methods of scene representation, including DL-based methods, which provided a great promise of DL-based methods for scene image representation. They categorised the range of methods into three broad categories. However, their study limits recent advances of DL-based methods in this domain.
(iv) Xie et al. [11] discussed the recent DL-based methods and traditional CV-based methods for scene representation, which not only carried out an in-depth study of each of them but also underscored the efficacy of DL-based methods against other methods for the scene image representation. However, their study has two main limitations. First, semantic approaches (e.g., SE-based methods) that have been gaining popularity recently are not included in their study. Second, their study lacks a comparative study of traditional CV-based methods, DL-based methods, and SE-based methods.

While looking into existing review works, we find the following gaps. First, the traditional CV-based methods are reviewed by most of the works, whereas the latest DL-based methods are not explored at their full potential. Second, the SE-based methods, which are recently introduced, also need in-depth analysis for their possible merits on scene image representations. Finally, the possible trend and research growth analysis are essential to show the possible research avenues but not available in the existing works.

To bridge the gaps in existing survey works, we study the recent and existing methods used in scene recognition and analyse them under their appropriate taxonomy using both qualitative and quantitative analysis. In addition, we present the ongoing research trends in scene image representation.

The main **contributions** in this paper are as follows:

(i) We perform a detailed review of the existing and recent scene image representation methods for classification using a comprehensive taxonomy.
(ii) We analyse the existing scene representation methods qualitatively and quantitatively. For quantitative analysis, we use a statistical approach, particularly box-plot analysis, across the performance measurein whereas, for qualitative analysis, we take the help of the pros/cons of methods.
(iii) Based on the pros and cons of the existing methods, we point out the potential directions of scene image representation and classification.
(iv) We reveal the trend and keyword growth analysis in the scene image representation area.

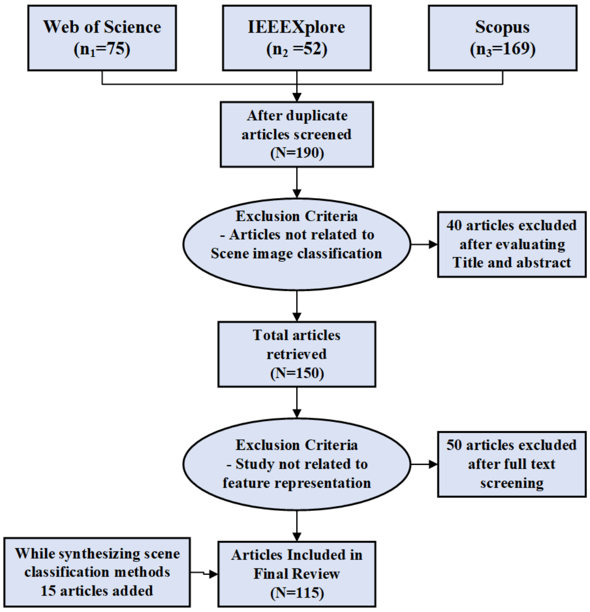


Fig. 1: Step-wise procedure to retrieve the articles reviewed in this survey.

The rest of the paper is organised as follows. Sec. 2 explains the process used to retrieve the papers for review. Similarly, Sec. 3 provides the basic concepts used in the scene representation, and Sec. 4 categorises the existing methods into three broad categories with their explanation. Sec. 5 explains the datasets used in the scene representation and details the comparative study of the existing methods and Sec. 6 discusses the overall methods and suggests the possible directions. Finally, Sec. 7 concludes the paper with final remarks.

2 Survey Method

In this section, we outline the procedure to retrieve the papers for review. We follow a systematic procedure to collect the papers for review. For this, we first search three popular databases: IEEE Xplore, Scopus, and Web of Science with the search string: "Scene Image OR Place" AND "Representation" AND "Classification". With this, we find 52, 169, and 75 articles with IEEE Xplore, Scopus, and Web of Science, respectively (Accessed date: 2022/11/10). After screening the title, abstract, author keywords, and full text, we end up collecting 100 articles. In addition to the searching method, we also collect 15 related articles using a snowballing technique. Last, a total of 115 articles are included for final review, including both scene representation methods and their related articles. The detailed pipeline of our survey method is presented in Fig. 1.

3 Background

Here, we explain the fundamental concepts, including both representation and classification algorithms, used in the literature mostly.

3.1 Representation algorithms

3.1.1 Scale Invariant Feature Transform (SIFT)

SIFT feature extraction algorithm, which was published in Lowe et al. [12], extracts the features based on the local sense of the image. This algorithm is mainly used for object recognition, gesture recognition, video tracking, etc.; however, it has also been used in scene representation problems [13]. It is a complex algorithm, which follows four steps to extract the descriptor: a) Scale-space detection, b) Key points localization, c) Orientation assignment, and d) Key points descriptor.

At first, to detect the key points in scale-space detection, multiple-scaled images are created and scale filtering is performed. For this, Laplacian of Gradient (LoG) could be used as a blob detection in each scale. However, since the LoG is a little bit costly, the Difference of Gaussian (DoG) is used in SIFT descriptor. The DoG is obtained by the difference of Gaussian blurring of an image with two differences σ, such as σ and $k\sigma$. Once the DoGs are achieved using such an approach, local maxima are found by searching the image with different scales and spaces. Local maxima are the potential key points of the corresponding image.

After the identification of potential key points in scale-space detection, the second step is to refine them for accurate results. For this, the Taylor series expansion algorithm [14] is used to get a more accurate location of local maxima in addition to the contrast threshold approach. With the help of the contrast threshold, we choose those extrema that have less than the threshold (e.g., 0.03), which can be chosen empirically. Furthermore, DoG exploits the edge information, which needs to be removed. Thus, the Harris corner detector is used to detect them and another threshold, called the edge threshold, is used to filter them out. With the help of such an approach, the extrema with low-intensity and edge key points are removed, thereby preserving only strong-intensity key points.

Next, the third step provides the in-variance to the extracted key points. In this step, orientation is assigned to each key point, where the neighborhood is considered into account around each key point depending on the scale, gradient, and direction. In this way, an orientation histogram is created with 36 bins covering 360 degrees. The highest peak of the histogram is taken and a peak below 80% is discarded.

Finally, the descriptor is created by taking the window of 16×16 neighborhood around the key points. Such a neighborhood is divided into 16 sub-blocks of 4×4, where for each sub-block, an orientation histogram of having 8 bins is constructed. This results in 128 bins in total for each key point. In this way, SIFT descriptor is created.

3.1.2 Histogram of Gradient (HoG)

HoG features also focus on the local sense, that is the gradient in the images. This concept was brought by Dalal et al. [15]. It was initially used to detect the objects in the image; however, it has been used in scene recognition problems these days [16]. To extract the HoG descriptors [17], we follow three steps: computation of gradient, orientation binning, and descriptor blocks.

First, the gradient values are calculated for an image. Specifically, this step utilizes filtering the color or intensity data of the image using two kernels such as [-1,0,1] and $[-1,0,1]^T$. Next, the histograms of cells are constructed. The structure of the cells can be either rectangular or radial and the histogram channels are spread over 0 to 180 or 0 to 360 degrees depending on the unsigned or signed gradient, respectively. Then, these histograms are normalized. Last, the HoG descriptor is obtained by the concatenation of all normalized cell histograms. Such blocks generally overlap, which means that each contributes more than once to form the descriptor.

3.1.3 Census Transform histogram (CENTRIST)-based features

The CENTRIST descriptor captures the structural detail of the image with the help of local structural detail. For this, spatial geometric information is utilized. To achieve such spatial information, it uses CT (Census Transform) values as its basic component. CT value is defined as the non-parametric local transform established to show the association between the intensity values [18]. To show the association in CT values, the intensity values are set to 0 if it is greater than the center value and set to 1 otherwise (Eq. (1)). Here, CT values (e.g, CT=224 for 20 in Eq. (1)) are calculated based on its 8 neighbouring intensity values. Finally, all the CT values are collected and constructed in the histogram to form the CENTRIST descriptor.

$$\begin{pmatrix} 10 & 20 & 30 \\ 10 & \mathbf{20} & 30 \\ 10 & 20 & 30 \end{pmatrix} \Rightarrow \begin{pmatrix} 1 & 1 & 0 \\ 1 & & 0 \\ 1 & 1 & 0 \end{pmatrix} \Rightarrow (11010110)_2 \Rightarrow 224 \tag{1}$$

Furthermore, the mCENTRIST [19] descriptor is the multi-channel CENTRIST descriptor, which is developed to overcome the weaknesses of CENTRIST. CENTRIST has mainly two weaknesses: first, it extracts the descriptor using a single channel; second, its descriptor size is larger. To overcome the weaknesses of CENTRIST, mCENTRIST uses complementary information using two or multiple channels, which improves the performance. Similarly, with the help of the Census Transform pyramid, they can reduce the size of the descriptor significantly.

3.1.4 Oriented Texture Curves

To achieve the OTC [20] descriptor, we need to perform three main steps. First, we need to sample the patches along the dense grid of the image. Next, each patch is represented by the curve, where each curve is based on a certain curve descriptor, that is texture-based and rotation sensitive. Note that for the texture-based descriptor, we use the HoG descriptor in the method. Last, such descriptor is concatenated and normalized to achieve the OTC descriptor.

3.1.5 Deep features

Deep features, which are the deep visual representation of the image, are extracted using various intermediate layers of deep learning model such as VGG16 [21]. Deep features achieved from different layers provide different kinds of information (e.g., foreground, background, etc.), which can be used to describe the various contents present in the image [21, 5, 22, 23]. Moreover, deep features represent the image at a higher order; therefore, it can discriminate such images more accurately than traditional computer vision-based descriptors such as SIFT, HoG, and so on.

3.1.6 Word embedding

Descriptors can also be achieved using the word embedding form from the pre-trained models [24, 25, 26]. Such descriptors, which are popular in Natural Language Processing (NLP) [27], have been used to extract the contextual information using tags/tokens representing the scene image [7]. There are basically three types of word embedding used in NLP tasks, which have also been used in image processing to capture contextual information. They are Word2Vec [24], GloVe [25], and fastText [26].

3.1.7 Sparse coding

Sparse coding yields the sparse representation of the input image based on the dictionary learning method. Based on the training images, a dictionary is constructed at first. Then, with the help of such a dictionary and its optimization, sparse representation to attain the final encoded features representing the image. This algorithm is popular in scene representation [28].

3.1.8 Bag of visual words

The bag of Visual Words (BoVW) encoding method is a slight variation of the bag of words (BoW) approach, which is quite popular in the Natural Language Processing (NLP) domain mostly. The BoVW method is invariant to scale and orientation, which is helpful to achieve better performance irrespective of the different resolutions and orientations of scene images. This method has been used widely in the computer vision domain nowadays [13]. To employ the BoVW in computer vision, the frequencies of visual words are considered, unlike the BoW approach.

3.1.9 Fisher vectors

To avoid the problem of sparsity and higher dimensionality problem in BoVW, the concept of Fisher vectors (FV) [29], which adopt the Fisher Kernel (the compact and dense representation), has been used. Specifically, the Fisher Vector (FV) is the general Fisher kernel, which is obtained by pooling local image features. For this, it stores the mean and covariance deviation vectors per component k of the Gaussian Mixture Model (GMM) in addition to each element of the local descriptor.

3.1.10 Locally-constrained Linear coding (LLC)

In LLC, each descriptor is projected to locality constraints using a local coordinate system and then, the projected coordinates are integrated using max-pooling operation, which results in the final representation [30]. This encoding is also popular to attain fixed-sized features for the scene image representation.

3.1.11 Principal Component Analysis

Principal Component Analysis (PCA) [31] has been used to reduce the dimension of the higher feature size. However, since it can provide fixed-sized features, it has also been used as an encoding algorithm. PCA extracts the orthogonal set of variables, which are called principal components (PCs). Based on those PCs, we achieve the reduced and fixed size of features. In the literature on scene image representation problems, this method has been used to reduce the deep feature size before the classification takes place [21].

3.1.12 Threshold-based histogram

This is an approach, where the fixed-sized features are constructed using the threshold operation to increment each bin of the histogram. Although this approach is computationally expensive, it can capture discriminating information. In scene representation, this approach has been used in SE-based algorithms to attain the feature vector representing the textual information [7].

3.2 Classification algorithms

After the representation of scene images, they are classified using either DL-based or traditional machine learning (ML)-based algorithms.

3.2.1 DL-based algorithms

DL-based algorithms learn the input data using different activities such as activation, convolution, pooling, and so on across several layers. In recent years, DL-based algorithms outperform traditional ML-based algorithms in most cases. This is because of their ability to learn several high-order information extracted from their intermediate layers. DL-based algorithms are divided into two categories: pre-trained and non-pre-trained. Pre-trained DL algorithms are open-access, which can be used as feature extractors for transfer learning or fine-tuning, whereas non-pre-trained models are user-defined DL algorithms, which are designed from scratch. The Softmax or Sigmoid layers are used for classification on top of those DL-based algorithms. Regarding the application of DL-based methods in the literature, it is noted that pre-trained DL algorithms have been mostly used for scene classification. For example, authors in [32, 33, 34] employed the pre-trained DL algorithms. The significant increment of performance from pre-trained DL algorithms due to transfer learning and fine-tuning is responsible for their widespread use in the literature.

3.2.2 Traditional ML-based algorithms

Traditional ML methods mostly rely on structured data and are simple to understand, implement, and interpret. They could work on limited data with limited hardware/resources, which makes it easier to deploy them in a resource-constrained setting. While looking at the literature on scene image classification, we notice that the Support Vector Machine [22, 35] is one of the most widely used traditional ML algorithms. This algorithm relies on the hyperplanes for the separability of images or data. It employs different kernels, including linear, polynomial, and radial basis functions. With the help of its complex kernels, it has been able to classify scene images. Similarly, researchers also used other algorithms such as nearest neighbour classifier [36], logistic regression classifier [21], and so on. The nearest neighbour algorithm classifies data based on the proximity of data. Similarly, the logistic regression (LR) algorithm employs the logistic function for the classification. It is interesting to see that traditional ML algorithms have been mostly used over deep features for scene image classification. This is because this approach helps improve the performance with the exploitation of both DL-based algorithms and traditional ML-based algorithms [21].

4 Taxonomy of scene image representation methods

In this section, we categorize the existing scene representation methods into three broad categories, which are traditional CV-based, DL-based, and SE-based methods (refer to Fig. 2 for the detailed taxonomy). The leaves of the taxonomy depict the algorithms for each method. Each method is explained in detail in the next subsections.

4.1 Traditional computer vision (CV)-based methods

Traditional computer vision-based methods [37, 38, 39, 20, 40] are based on the basic components of the image such as colours, pixels, lines, and shapes. The use of such basic components helps us understand how images are constructed and based on such patterns, we can represent them easily for several tasks such as classification, clustering, recognition, and prediction. The high-level flow of traditional computer vision-based methods for scene image representation and classification is presented in Fig. 3, which includes three steps: feature extraction, feature encoding, and classification.

Most popular traditional image representation methods are based on Generalized Search Trees (Gist) [41, 37], Gist-Color [37], CENsus TRansform hISTogram (CENTRIST) [39],multi-channel (mCENTRIST) [19], Scale-Invariant Feature Transform (SIFT) [38], Histogram of gradient(HoG) [15], Oriented Texture Curves (OTC) [20], Object bank representation(OBR) [42, 43], SPM [13], Reconfigurable BoW (RBoW) [44], Bag of Parts (BoP) [45], Important Spatial Pooling Region (ISPR) [46], etc. Among these techniques, the popular method such as Gist extracts the features from local details such as color, pixels, and orientation of images [37, 47, 48, 42, 44, 45, 46, 49, 50]. Therefore, they are limited to dealing with high variations in the local image features. Furthermore, the OTC [20] method extracts

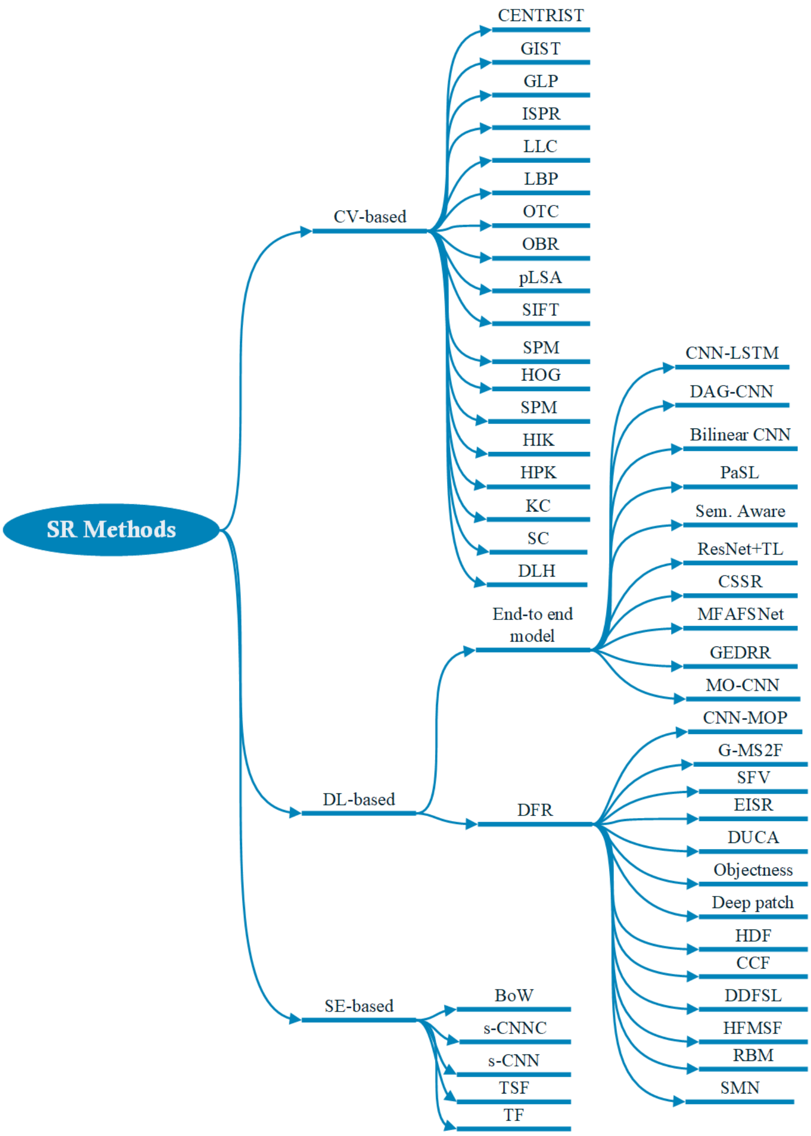


Fig. 2: Taxonomy of existing scene image representation methods

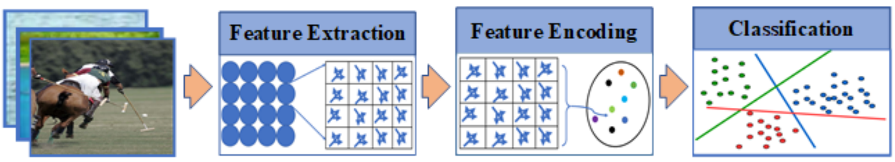


Fig. 3: CV-based scene representation pipeline for classification

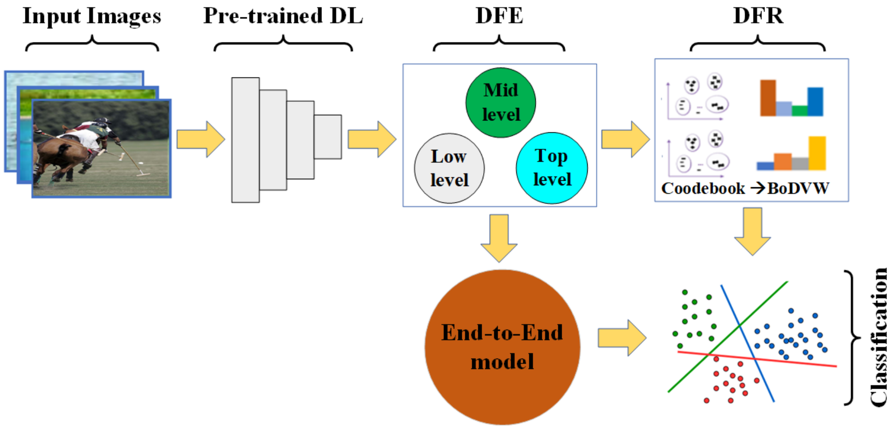


Fig. 4: DL-based scene representation pipeline for classification

the image features based on the colour variation of various patches in images, keeping in mind that these features are suitable to represent the texture images, not much pertinent to scene images. However, Spatial Pyramid Matching (SPM) [13] employs SIFT, which are multi-scale and rotation-invariant local features. Going forward, SPM first slices the images and then extract image feature based on those spatial regions of the image. The extracted features of each region are represented as a Bag of Visual Words (BoVW) of SIFT descriptors. Even though this method captures more semantic regions than other methods of the scene image to some extent, they are still not suitable to represent complex scene images requiring high-level information such as object and foreground/background information for discriminability.

4.2 Deep learning (DL)-based methods

Deep learning models, which are a composition of multiple artificial neural networks [51], have provided a breakthrough performance in various domains such as text classification [52, 27], health informatics [53] and computer vision [23, 54]. Among three different methods, DL-based methods are most popular today to represent and classify scene images. The high-level diagram of DL-based methods is presented in Fig. 4, which includes deep feature extraction (DFE) using pre-trained models (e.g., low-level, mid-level, and high-level), deep feature representation by encoding approach (e.g., a bag of words, fisher vector, etc.), and classification. Besides, some DL methods prefer training in an end-to-end fashion after the deep feature extraction (DFE) step for the classification.

There are two approaches/techniques (uni-modal and multi-modal) preferred by most of the DL-based methods for scene image representation and classification. First, there are some works in scene representation and classification that use uni-modal pre-trained deep models such as ResNet152 [55], VGG-Net [56, 57, 32], AlexNet [58], GoogleLeNet [59], and HDF [23]. For example, authors in [60] extracted features from VGG-Net pre-trained on hybrid datasets (ImageNet [61] and Places [62]) using Caffe [63] platform. They used fully connected layers (FC), which resulted in a feature size of $4,096$-D for each scale of the image to achieve orderless multi-scale pooling features. The final feature size of their method is higher as the number of scales increases in their experiment. Their method outperforms the single-scaled features though their method has a higher dimensional feature size. Similarly, authors in [64] used features from VGG-Net pre-trained on ImageNet [61] and extracted the high-level feature from the FC-layers after a fine-tuning operation. These features were fed into the Naive Bayes non-linear algorithm [65] for the classification. The performance of their method is promising; however, their method requires a massive dataset for fine-tuning operations, which could limit its applicability in real time. Furthermore, authors in [66] utilized three classification layers of fine-tuned GoogleNet [59] model, where they extracted the deep features in the form of probabilities and then performed the features fusion to achieve the results. Although their method outperforms several existing methods in the literature, it requires large datasets for fine-tuning coupled with an arduous hyper-parameter tuning operation to learn the highly separable features.

Furthermore, some studies improved the separability of scene images by extracting the mid-level features from the pre-trained deep learning models. For instance, Zhang et al. [67] randomly cropped the image into multiple patches and extracted the visual features from each of them using the AlexNet [58] model. Then, these features were used to design the codebook of size $1,000$-D for the sparse coding technique to extract the relevant features. Later on, they concatenated the sparse coded features with the tag-based features to get the final features for the classification. Because of highly discriminating features from both deep features and sparse coded features, their method imparts a significant boost in performance compared to the existing methods. However, their work possesses two main limitations: a) the chance of feature repetition as the patches are selected randomly; and b) higher feature size. In addition, bag of surrogate parts (BoSP) features were proposed by Guo et al. [68] based on the two higher pooling layers– 4^{th} and 5^{th} of the VGG16 model [56] pre-trained on ImageNet [61]. However, their method only captures the foreground information as they employed the VGG-16 model pre-trained on ImageNet. As a result, it lacks the background information, which is one of the important clues required to better discriminate the complex scene images having higher inter-class similarity and intra-class dissimilarity. Additionally, authors in [69] compared four different CNN models such as AlexNet [58], ResNet152 [55], VGG-16 [56], and GoogleLeNet [59] pre-trained on ImageNet and Places datasets for scene image classification using semantic multinomial representation (SMN) approach, where they utilized pre-trained models available for Caffe [63] model zoo without fully connected layers and fine-tuning operation. This is one of the recent methods used in scene image representation and classification, which has shown great promise against the existing methods.

Second, a few works proposed to use multi-modal deep features to represent the scene image for classification. For instance, Sun et al. [96] used three models: YOLOV2 [97], HybridDNN [96], and VGG-16 to represent the scene images. Here, the global appearance feature (GAF) from the second-last layer of VGG-16, CFA feature from the hybrid DNN and spatial layout maintained object semantics feature (SOSF) from the YOLOV2 models were concatenated to represent the scene image. The resultant features were trained using the SVM classifier. Moreover, Bai et al. [32] proposed a multi-modal architecture utilizing both CNN and Long Short

Table 2: Dataset description used in scene image representation and classification.

Dataset	Type	Highlights	Ref.
MIT-67	RGB	Complex scene images	[20, 46, 19, 70, 60, 71, 66, 72, 62, 67, 73, 74, 75, 76, 77, 32, 78, 23, 79, 80, 21, 22, 81, 82, 83, 84, 7, 28, 21, 22]
Scene-15	Grayscale	Indoor-outdoor images	[37, 13, 85, 86, 87, 39, 20, 88, 46, 70, 71, 66, 67, 75, 76, 23, 79, 22, 84, 83, 7, 16, 89, 28, 22]
Event-8	RGB	Sport events related images	[37, 39, 88, 46, 19, 67, 75, 23, 79, 22, 90, 84, 7, 16, 22]
SUN-397	RGB	Complex indoor/outdoor scene images	[20, 60, 71, 66, 72, 62, 73, 74, 76, 32, 78, 80, 21, 81, 82, 28]
Caltech-256	RGB	Natural and artificial objects in a diverse setting	[91, 87, 92, 93]
NYU-V1	RGB-Depth	Indoor images with RGB and depth information	[94, 95]

Term Memory (LSTM) model for the scene image classification. The LSTM model was used on top of CNNs. In their proposal, each image slice feature was extracted from VGG-16 [56] pre-trained on Places [62] and then, fed into the LSTM model. Since the deep learning model pre-trained model on the Places dataset gives the background information and LSTM captures the sequence information of image slices, their model outperforms several other previous methods, including traditional CV-based methods and several DL-based methods. Furthermore, Liu et al. [98] proposed to use the CNN features and euclidean distance approach, which improved the performance on both MIT-67 and Scene-15 datasets. Furthermore, considering the popularity of metric learning and local manifold preservation, authors in [34] proposed a novel approach called, a joint global metric learning and local manifold preservation (JGML-LMP), which provided a significant boost in the classification performance.

A few works on scene image classification used the whole-part feature extraction approach using both foreground and background information. For instance, the whole- and part-level feature extraction approach was proposed by Sitaula et al. [23] to represent the scene images. In their method, they utilized pre-trained VGG model on both ImageNet [61] and Places [62] to capture both foreground and background information for each input scene image. Since their method does not consider contextual information, it still provides a limited performance while dealing with complex scene images having a higher inter-class similarity.Authors in [99] also employed the object-centric and place-centric information or features to classify the indoor images.

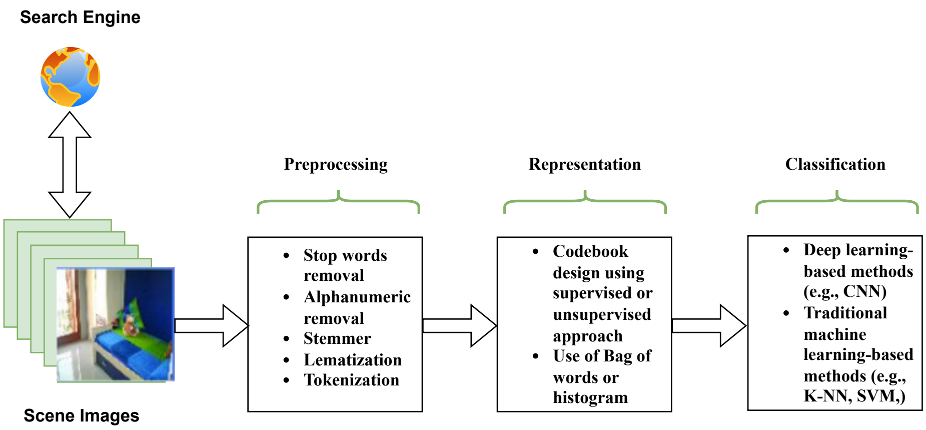


Fig. 5: Search engine (SE)-based scene representation pipeline for classification.

4.3 Search engine (SE)-based methods

The visual information achieved from either traditional CV-based or DL-based methods is not sufficient to represent the complex scene images because they also require contextual information (e.g., non-visual information such as tags, tokens, and annotation) for their accurate separability. There are very few works [67, 83, 7], which extract contextual information using a search engine, for the representation of scene images in the literature. These methods are considered SE-based methods. While the extraction of features related to scene images using search engines is an arduous process, it still has an immense potential to differentiate complex scene images due to the presence of human annotations/descriptions for similar images on the web. The high-level diagram of SE-based methods is presented in Fig. 5, which comprises three steps: preprocessing (e.g., stop words removal, stemmer, etc.), representation (e.g., codebook, histogram, etc.) and classification.

Under the SE-based methods, authors in [67] collated the annotations/tags of top 50 visually similar searched images for the phrased input query image on the web. The collated tags were preprocessed and classified in an end-to-end fashion. The main limitation of their work is the higher feature size incurred by the bag of words on raw tags, which could be minimized by using the filter bank. Later on, the idea of filter banks to minimize the feature size was established by Wang et al. [83], where they proposed the task-generic filter banks using the pre-defined category names to filter out the outlier tags to some extent. For the pre-defined category names, they borrowed them from the ImageNet [61] and Places [62] datasets. However, their method still lacks domain-specific keywords/tags related to scene images, which could lead to out-of-vocabulary problems. As a result, it creates an accumulation of unnecessary tags in the filter banks. This, in the end, could ultimately degrade the classification accuracy. Given such limitations, Sitaula et al. [7] constructed the domain-specific filter bank based on the training data. Their domain-specific filter bank not only helped minimize the vocabulary problems but also improved the overall classification performance of scene images as they were able to capture more semantic information. By and large, the contextual information captured from the web can provide important clues to discriminate complex scene images having both inter-class similarity and intra-class dissimilarity [83, 7].

5 Datasets

Although several datasets, including both smaller and larger ones, have been used in the literature for scene representation and classification, we list and explain the commonly-used larger scene image datasets in this study. There are commonly six benchmark datasets (MIT-67 [47], Scene-15 [100], Event-8 [101], SUN-397 [102], Caltech-256 [91], and NYU-V1 [94]), which have been used frequently in the literature.

MIT-67 [47] contains $15,620$ images (67 categories), where each category contains at least 100 images. There is a standard protocol [47] of train/test protocol to be used in the experiments. According to the protocol, 80 images per category are taken as the training split, whereas 20 images per category are taken as the testing split.

Scene-15 [100] contains $4,485$ images (15 categories), where each category contains at least 200 images. There is no standard train/test protocol defined to use this dataset. However, researchers use 100 images per category as training and the rest of the images as testing split. The experiment is repeated for 10 runs to report the average accuracy.

Event-8 [101] contains $1,579$ images (8 categories), where each category contains at least 137 images. There is no standard train/test split ratio to use this dataset; however, researchers randomly select 120 images per category and divide 70 images as training and 60 images per category as a testing split. The experiments are conducted for 10 runs to note the average accuracy.

SUN-397 [102] contains $108,754$ images (397 categories), where each category contains at least 100 images. This dataset provides standard 10 sets of train/test protocol [102] to be used in the experiments, where each split contains 50 images/category as training and 50 images/category as testing. The average of 10 runs is used to report the accuracy.

Caltech-256 [91] contains $30,607$ images(256 object categories). It consists of images of various natural and artificial objects in diverse settings. The minimum number of images in each category is 80.

NYU-V1 [94] consists of $2,347$ labeled frames having 7 different classes. The images were collected from a wide range of domains, where the background was changing from one to another with RGB and depth cameras from the Microsoft Kinect. Given that scene images in this dataset contain several objects and their associations, this dataset is one of the most challenging datasets for scene image classification. Summary details of all of these datasets are mentioned in Table 2.

6 Discussion

Here, we discuss the research works carried out in scene representation and classification using quantitative (e.g., performance metrics) and qualitative analysis (e.g., pros/cons).

Table 3: Comparative study of state-of-the-art methods using classification accuracy (%) on scene datasets under CV-based methods. The symbol − represents the no published accuracy.

Approach	Scene-15	Event-8	MIT-67	SUN-397
Gist-color [37]	69.5	70.7	-	-
SPM [13]	72.2	-	-	-
pLSA [85]	72.7	-	-	-
Semantic Theme [86]	72.2	-	-	-
Kernel Codebook [87]	76.7	-	-	-
CENTRIST [39]	84.9	78.5	-	-
OTC [20]	84.3	-	47.3	34.5
S^3R [88]	83.7	40.1	-	-
ISPR [46]	85.0	89.5	50.1	-
WSR-EC [70]	81.5	-	38.6	-
mCENTRIST [19]	86.5	44.6	-	-
Xie et al. [16]	83.3	84.8	-	
Ali et al. [89]	90.4	-	-	
HIK[103]	-	-	40.19	-
HPK [104]	-	-	-	-
HPK [104]	-	-	-	-
HILLC [105]	86.3	85.0	-	-
CS-PSL [92]	-	-	52.5	-
OBR [43]	88.8	86.0	32.3	-
3-DLH [36]	-	84.9	-	-
LLC [30]	83.2	-	-	-
PFE [106]	84.2	-	-	-
SIFT[94]	-	-	-	-
W-LBP[107]	85.1	86.2	-	-
GPHOG [40]	-	-	-	-
Spatial LBP [35]	80.9	71.7	-	-
BoW-LBP [36]	80.7	87.7	-	-

6.1 Quantitative analysis

For the quantitative analysis of research articles published in the literature, we summarise the performance using box plots, which impart the statistical information of classification performance, as shown in Fig. 6. (Note that we draw boxplots based on the performance of three different scene representation methods (DL-based, CV-based and SE-based) achieved from the corresponding Tables 3, 4 and 5 on four datasets (Figs. 6(a), 6(b), 6(c) and 6 (d), respectively.)

Here, we analyze the performance, particularly the reported accuracies of three or two different methods on four datasets. Since the search engine (SE)-based methods only consider three datasets (Scene-15, Event-8, and MIT-67) in the literature, we present the results on only such three datasets, whereas, for the other two methods (DL-based and CV-based), we present the results on four datasets (Scene-15, Event-8, MIT-67, and SUN-397).

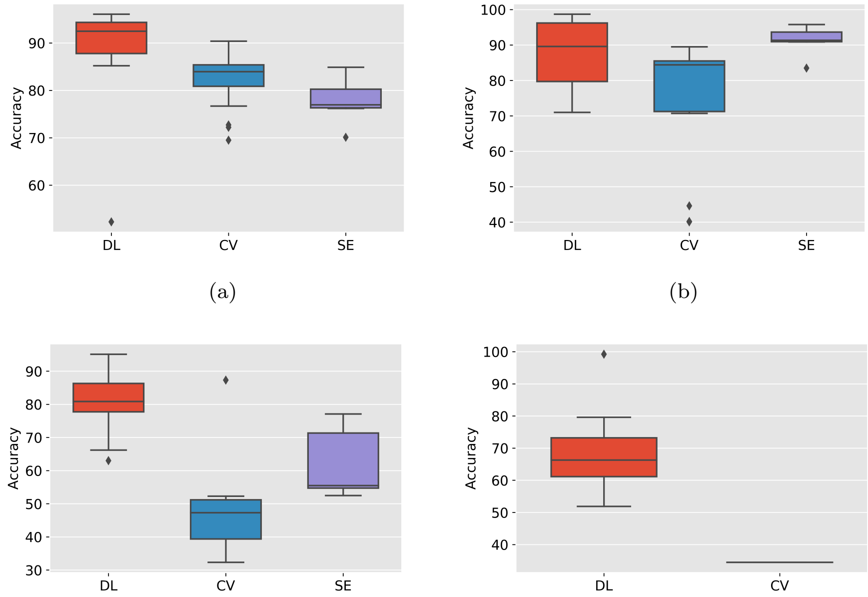
While comparing the performance of three different kinds of methods on four datasets, we notice that DL-based methods outperform other remaining methods in all datasets. For example, on the Scene-15 dataset, DL-based methods provide the highest accuracy mostly (maximum and minimum of 98.7% from RBM [113], and 85.2% from ResNet+TL [109], respectively) compared to the traditional CV-based methods that has below 85% accuracy mostly except Ali et al. [89] with

Table 4: Comparative study of state-of-the-art methods using classification accuracy (%) on four scene datasets under DL-based methods. The symbol − represents the no published accuracy.

Approach	Scene-15	Event-8	MIT-67	SUN-397
CNN-MOP [60]	-	-	68.8	51.9
DAG-CNN [71]	92.9	-	77.5	56.2
G-MS2F [66]	92.9	-	79.6	64.0
SFV+Places [72]	-	-	79.0	61.7
VGG [62]	91.72	95.17	79.7	63.2
EISR [67]	92.1	89.6	66.2	-
VSAD [73]	-	-	86.2	73.0
LS-DHM [74]	-	-	83.7	67.5
DUCA [75]	94.5	98.7	71.8	-
Nascimento et al. [28]	95.7	-	87.2	71.0
Objectness [76]	95.8	-	86.7	73.4
Bilinear-CNN [77]	-	-	79.0	-
Deep patch [78]	-	-	79.6	57.4
HDF [23]	93.9	96.2	82.0	-
Sorkhi et al. [79]	95.1	99.2	73.6	-
PaSL [80]	-	-	88.0	74.0
Semantic-Aware [81]	-	-	87.1	74.0
LASC [82]	-	-	81.7	64.3
FBH [21]	-	-	82.3	66.3
CCF [22]	95.4	98.1	87.3	-
DDSFL [108]	52.2	86.9	84.4	-
ResNet+TL[109]	85.2	-	94.0	-
HFMSF[110]	97.8	-	-	-
CNN-LSTM[32]	-	-	80.5	63.0
ABR [111]	91.9	96.2	68.3	-
CSSR [112]	-	-	77.8	57.3
RBM [113]	98.7	-	-	-
SOSF+CFA+GAF [96]	-	-	89.5	78.9
DeepFeature [114]	-	94.8	72.3	-
SMN [69]	-	-	84.4	66.8
RVF [115]	-	-	80.0	60.6
MFAFSNet [116]	-	-	88.0	72.4
GEDRR [117]	96.0	-	87.7	73.5
MetaObject +CNN [118]	-	-	78.9	58.1
JGML-LMP[119]	96.0	99.0	87.5	73.2
Liu et al. [34]	96.4	-	81.6	-
Selective CNN [34]	-	-	88.4	-

Table 5: Comparative study of state-of-the-art methods using classification accuracy (%) on four scene datasets under SE-based methods. There are no reported accuracies on SUN-365 dataset using such methods.

Approach	Scene-15	Event-8	MIT-67
BOW [83]	70.1	83.5	52.5
s-CNN(max) [83]	76.2	90.9	54.6
s-CNN(avg) [83]	76.7	91.2	55.1
s-CNNC(max) [83]	77.2	91.5	55.9
TSF [7]	81.3	94.4	76.5
TF [22]	84.9	95.8	77.1



(a) (b)

(c) (d)

Fig. 6: Box-plot visualization of summary accuracy (%) achieved by three different methods on four most popular scene image datasets: (a) Scene-15, (b) Event-8, (c) MIT-67, and (d) SUN-397. Note that DL, CV, and SE represent DL-based, CV-based, and SE-based methods. Note that there is no reported accuracy for SE-based methods on the SUN-397 dataset.

90.4% accuracy. The reason for such performance surge while using DL-based methods is because of the highly discriminating feature extraction abilities from different intermediate layers of DL methods. Notably, deep features could provide more information related to scene images, including foreground, background, and hybrid. The presence of all three kinds of information helps discriminate the complex scene images more accurately. However, traditional CV-based methods are not sufficient to capture such information, which as a result fails to discriminate the complex scene images during classification. Also, the recent works using the search engine (SE)-based methods on three datasets (Scene-15, Event-8, and MIT-67) show that SE-based methods could capture complementary contextual information, which is difficult to achieve from the visual information achieved from the traditional CV-based and DL-based methods, for the scene images to represent them during classification. Interestingly, it can outperform the traditional CV-based methods and is comparable to DL-based methods during scene image representation and classification. For example, SE-based methods on the Event-8 dataset (6(b)) provide an accuracy of over 90%, whereas the traditional CV-based methods and DL-based methods provide an accuracy below 90% and over 90%, respectively. This encouraging classification performance shows the efficacy of SE-based methods for scene image representation.

While comparing the performance throughout the four widely popular datasets (Scene-15, Event-8, MIT-67, and SUN-397) reported in Fig. 6, we observe that SUN-397 is the most challenging dataset for which the state-of-the-art methods have produced the least performance compared to the other three datasets (Scene-15, Event-8, and MIT-67). Also, there is no reported classification accuracy for SE-based methods for this dataset. Furthermore, the accuracy of SUN-397 remains between around 71% and 35% in the classification. We believe that this is the most challenging dataset compared to other datasets, both in terms of complexities (higher inter-class similarity and intra-class dissimilarity) and categories (higher number of challenging classes). Similarly, we observe that the MIT-67 dataset is the second-most challenging dataset in terms of performance, which has a maximum performance of around 97% by DL-based methods and a minimum performance of around 40% by CV-based methods. Although this dataset has only 67 categories compared to SUN-397 (397 categories), it is still a challenging dataset with a similar level of complexity to SUN-397 for scene image representation and classification. Compared to the SUN-397 and MIT-67 datasets, two other datasets (Scene-15 and Event-8) are relatively less challenging and have produced the most prominent classification performance (Scene-15 has the maximum and minimum accuracy of over 98% by DL-based methods and over 76%, by SE-based methods respectively, whereas the Event-8 has the maximum and minimum accuracy of over 95% by DL-based methods and over 70% by CV-based methods, respectively). The reason for such a significant boost in performance is attributed to the distinguishable scene images (lower inter-class similarity and intra-class dissimilarity) present in them.

To sum up, the DL-based methods outperform both the traditional CV-based method and SE-based methods in most cases. This infers that visual content information of the scene images provided by the DL-based methods is more discriminating than others to distinguish ambiguous and complex scene images. Recently, the SE-based methods have shown some promise in scene image representation by providing some important contextual clues, which are attained using human perception and knowledge available on the internet.

6.2 Qualitative analysis

Here, we analyse each of the three methods (CV-based, DL-based, and SE-based) based on their advantages and shortcomings, which are obtained in terms of their viability.

Regarding CV-based methods, they have four major merits. First, feature extraction is well-established and easier to implement. For example, we can achieve the features based on the traditional CV-based methods such as SIFT (Scale Invariant Feature Transform) and HoG (Histogram of Gradient) with a few lines of code. Second, they have a higher performance with fine-grained and non-ambiguous images (no inter-class similarity and intra-class dissimilarity). With the help of basic information of scene images such as pixels, lines, and arc details, it is easy to distinguish the non-complex images (e.g., fine-grained, texture, non-ambiguous, etc.) during classification. Third, CV-based methods are less complex compared to other methods because they do not require arduous training activities to achieve the discriminating features of the input image. Fourth, we do not require a domain-specific knowledge to implement them. For example, we can apply the same SIFT algorithm for both scene images and biomedical images to represent them. In contrast, CV-based methods have two major demerits. First, they have a lower classification performance for complex scene images having higher inter-class similarity and intra-class dissimilarity. This is because complex scene images require a higher level of information (e.g., object), which is difficult to acquire by CV-based methods. Second, given that there are several kinds of features achieved from the CV-based methods, it is very difficult to choose the most discriminating and useful features corresponding to the study.

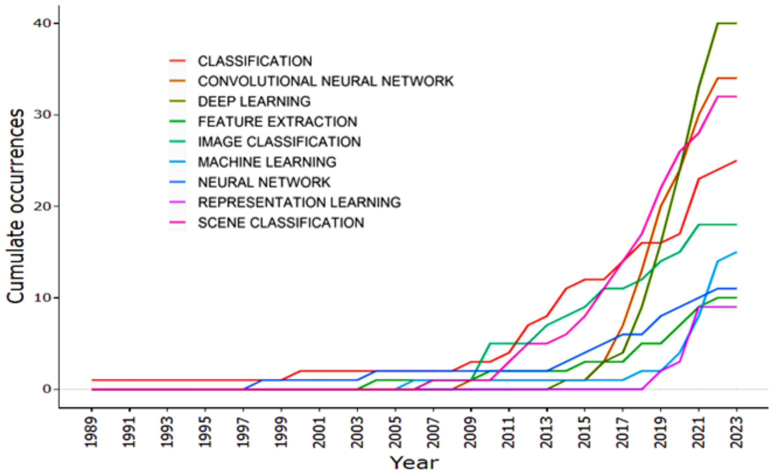
For the DL-based methods, they have two major merits. First, they have a higher classification performance on complex images compared to CV-based methods. This is because they can extract the high-level information (e.g., object) present in the scene image. Second, DL-based methods are flexible. That is, the DL models can be re-trained using custom datasets unlike the CV-based methods to make them domain-specific. Nevertheless, DL-based methods have three major demerits. First, they are heavy-weight in most cases compared to CV-based methods. The DL-based methods are very difficult to deploy in the edge computing environment as they require heavily trained weight files to achieve promising accuracy. Second, the training and re-training processes of DL-based models are labor-intensive as they are prone to over-fitting and under-fitting problems. Third, although they have higher accuracy compared to others, they are, in most cases, poor in interpretability and explainability.

The SE-based methods have two major merits. First, they can capture contextual information with the help of human knowledge, which is complementary information to visual features. Second, the combination of contextual information with visual information could overcome the limitations of each individual. In contrast, they have two major demerits. First, they are computationally infeasible to capture the information via search engines if we have a massive number of images because search engines have a restriction on the number of query inputs for searching. Second, while selecting the tokens or textual information online, it is very difficult to select the most important information from the annotations/tags as we encounter numerous significant pieces of information. Since the current works focus on top-k images for annotations/tags, they could end up missing some important information present beyond k images.

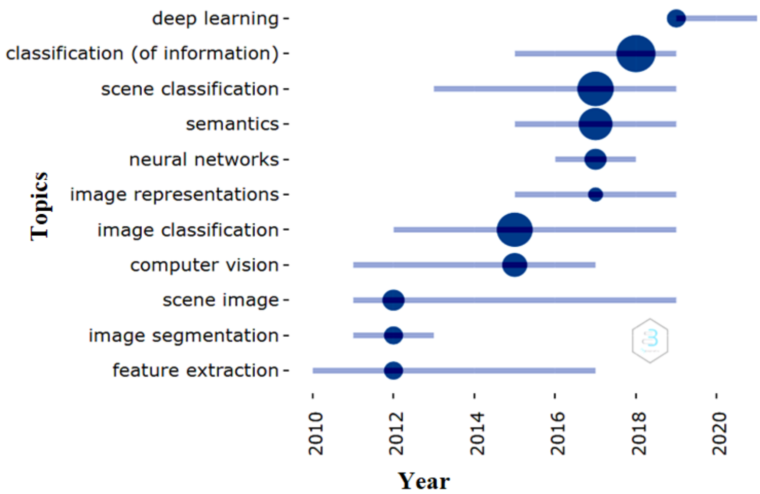
6.3 Research trend analysis

Here, we analyze the research direction of scene representation based on the cumulative occurrence of keywords and time duration across different years using a Line graph and Forest plot [120], respectively, which are presented in Fig. 7. The frequently-used keywords help understand the research direction in scene analytics because they not only provide the frequency but also their inception and current state. In this study, such keywords have been picked by the Forest plot automatically based on their importance.

While looking at Fig. 7 in terms of topic occurrence, we observe that the cumulative topic occurrence has been increasing from 1996 to this date. There have



(a)



(b)

Fig. 7: Author's keyword growth during last decades

been several topics popular in scene image representation such as 'classification (of information)', 'computer vision, 'deep learning, and 'semantic'. Among them, it is noted that 'classification(of information)' is the most popular topic, which has been sharply increasing in recent years. In addition, some other topics such as 'scene classification', and 'feature extraction' are also following similar kinds of patterns, whereas other topics such as image segmentation and scene classification are increasing at a slower rate. We believe that this trend makes sense because basic works related to scene image representation have already been done such as 'scene classification' and 'feature extraction. The current need is to build robust AI models with higher performance. Overall, the research trend of different topics in scene images has been in the upward direction with the predominant use of DL-based methods.

While analyzing the keyword topics' popularity in terms of time duration at Fig. 7, we notice that different topics have different time duration for their popularity level. For example, from 2010 to 2017, most of the research works in scene representation were focused on feature extraction and it was most popular in 2012. We believe that this is because feature extraction is the foundation work of scene image representation. It is seen that most of the research topics in scene image representation such as 'semantics', 'neural networks, 'scene classification', and 'classification' are quite popular after 2017. In recent days, particularly after 2019, 'deep learning has become a prominent topic, which is because of the groundbreaking classification performance produced by them. To this end, the popularity of different keywords in different years reveals the different levels of research in scene representation and classification.

7 Conclusion and future works

In this paper, we have reviewed the research works carried out in the scene image representation area for classification and categorised them into three broad groups: CV-based, DL-based, and SE-based methods. This categorisation and analysis (both qualitative and quantitative) reveal that the DL-based methods outperform the remaining two methods in terms of classification accuracy in most cases, whereas the SE-based methods remain the potential research direction in the future. We also find that the DL-based methods have been frequently used in recent years using a transfer learning approach for performance improvement, whereas the SE-based methods, which are on the rise, have shown difficulty because of search engines although they have a great promise. We also underline that the combination or fusion of the DL-based methods with other methods enhances the classification performance significantly, which is because of the rich information obtained from multiple sources during image representation. In addition, we find that scene representation research works (e.g., feature extraction, representation learning, scene classification, etc.) are on the rise in recent years.

Furthermore, we notice that the usability of the method for the scene image representation is dependent on the requirements. If the requirement is on a performance issue, it is inevitable to use the DL-based methods as they provide a groundbreaking performance; however, they require higher computational and space requirements. As such, we encourage building domain-specific lightweight pre-trained DL models to be used in the future. Given that our current study does not include the application of domain-specific lightweight DL models on scene image analytics followed by their trend analysis, we believe that it could be an interesting survey study in the future.

8 Data availability

All data are publicly available.

9 Abbreviations

The list of abbreviations used in our study is presented in Table 6.

Abbrv.	Full form
ABR	Attribute-Based high-level image Representation
BSRC	Block Sparse Representation Based Classifier
CCF	Content Context Features
CFA	Contextual Features in Appearance
CSSR	Category-Specific Salient Region
CS-PSL	class-specific pooling shapes Learning
DDSFL	Deep Discriminative and Shareable Feature Learning
DoG	Difference of Gaussian
DAG-CNN	Directed Acyclic graph-Convolution Neural Network
DUCA	Deep Un-structured Convolutional Activation
EISR	Explicitly and Implicitly Semantic Representations
FBH	Foreground background hybrid features
GAF	Global Appearance Feature
GEDRR	Global and Graph Encoded Local Discriminative Region Representation
Gist	Generalized Search Trees
GPHOG	Gabor Pyramid of Histograms of Oriented Gradients
G-MS2F	GoogLeNet-based Multi-Stage Feature Fusion
GMM	Gaussian Mixture Model
HDF	Hybrid deep features
HFMSF	Handcrafted Features with Deep Multi-stage Features
HIK	Histogram Intersection Kernel
HILLC	Histogram Intersection-Locally-constrained Linear coding
HPK	Hybrid Pyramid Kernel
ISPR	Important Spatial Pooling Region
IoT	Internet of Things
LoG	Laplacian of Gradient
LASC	Locality-constrained Affine Subspace Coding
LS-DHM	Locally Supervised Deep Hybrid Model
LSTM	Long short-term memory
MFAFSNet	Mixture of Factor Analyzers-Fisher Score Network
MOP	Multiscale orderless pooling
OTC	Oriented Texture Curves
OBR	Object Based Representation
pLSA	probabilistic Latent Semantic Analysis
PFE	Pooled Feature Extraction
RBM	Restricted Boltzman Machine
RVF	Reduced Virtual Features
SC	Sparse coding
SIFT	Scale-Invariant Feature Transform
SOSF	Spatial-layout maintained Object Semantics Features
SPM	Spatial Pyramid Matching
SMN	semantic Multinomial Network
S^3R	Sub-semantic space
SFV	Semantic Fisher Vectors
TSF	Tag-based semantic features
TF	Tag-based features
VGG	Visual Geometry Group
VSAD	Vector of Semantically Aggregating Descriptor
W-LBP	Wigner-based Local Binary Patterns
WSR-EC	Weak semantic image representation- Example classifier
3-DLH	3-Dimensional LBP-HaarHOG

Table 6: List of abbreviations used in this study

10 Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Sitaula C, Xiang Y, Zhang Y, Lu X, Aryal S (2019) Indoor image representation by high-level semantic features. IEEE Access 7:84,967–84,979
2. Shadman Roodposhti M, Aryal J, Lucieer A, Bryan BA (2019) Uncertainty assessment of hyperspectral image classification: Deep learning vs. random forest. Entropy 21(1):78
3. Neupane B, Horanont T, Aryal J (2021) Deep learning-based semantic segmentation of urban features in satellite images: A review and meta-analysis. Remote Sensing 13(4):808
4. Dutta R, Aryal J, Das A, Kirkpatrick JB (2013) Deep cognitive imaging systems enable estimation of continental-scale fire incidence from climate data. Scientific reports 3(1):1–4
5. Sitaula C, Xiang Y, Aryal S, Lu X (2019) Unsupervised deep features for privacy image classification. In: Proc. Pacific-Rim Symposium on Image and Video Technology (PSIVT), pp 404–415
6. McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics 5(4):115–133
7. Sitaula C, Xiang Y, Basnet A, Aryal S, Lu X (2019) Tag-based semantic features for scene image classification. In: Proc. Int. Conf. on Neural Inf. Process. (ICONIP), pp 90–102
8. Wei X, Phung SL, Bouzerdoum A (2016) Visual descriptors for scene categorization: experimental evaluation. Artif Intell Rev 45(3):333–368
9. Anu E, Anu K (2016) A survey on scene recognition. Int J Sci Eng Technol Res(IJSETR) 5:64–68
10. Singh V, Girish D, Ralescu A (2017) Image understanding-a brief review of scene classification and recognition. In: Proc. Modern Artificial Intelligence and Cognitive Science (MAICS), pp 85–91
11. Xie L, Lee F, Liu L, Kotani K, Chen Q (2020) Scene recognition: a comprehensive survey. Pattern Recognit p 107205
12. Lowe DG (1999) Object recognition from local scale-invariant features. In: Proc. Int. Conf. Comput. Vis. (ICCV), vol 2, pp 1150–1157
13. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 2169–2178
14. Moller T, Machiraju R, Mueller K, Yagel R (1997) Evaluation and design of filters using a taylor series expansion. IEEE transactions on Visualization and Computer Graphics 3(2):184–199
15. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 886–893
16. Xie L, Lee F, Liu L, Yin Z, Yan Y, Wang W, Zhao J, Chen Q (2018) Improved spatial pyramid matching for scene recognition. Pattern Recognition 82:118–129

17. Ayalew AM, Salau AO, Abeje BT, Enyew B (2022) Detection and classification of covid-19 disease from x-ray images using convolutional neural networks and histogram of oriented gradients. Biomedical Signal Processing and Control 74:103,530
18. Zabih R, Woodfill J (1994) Non-parametric local transforms for computing visual correspondence. In: Proc. Euro. Conf. Comput. Vis. (ECCV), pp 151–158
19. Xiao Y, Wu J, Yuan J (2014) mcentrist: a multi-channel feature generation mechanism for scene categorization. IEEE Trans Image Process 23(2):823–836
20. Margolin R, Zelnik-Manor L, Tal A (2014) OTC: A novel local descriptor for scene classification. In: Proc. Eur. Conf. Comput. Vis. (ECCV), pp 377–391
21. Sitaula C, Xiang Y, Aryal S, Lu X (2021) Scene image representation by foreground, background and hybrid features. Expert Systems with Applications p 115285
22. Sitaula C, Aryal S, Xiang Y, Basnet A, Lu X (2021) Content and context features for scene image representation. Knowledge-Based Systems p 107470
23. Sitaula C, Xiang Y, Basnet A, Aryal S, Lu X (2020) HDF: Hybrid deep features for scene image representation. In: Proc. Int. Joint Conf. on Neural Networks (IJCNN)
24. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781
25. Pennington J, Socher R, Manning C (2014) Glove: Global vectors for word representation. In: Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP), pp 1532–1543
26. Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans of the Association for Computational Linguistics 5:135–146
27. Shahi TB, Sitaula C (2021) Natural language processing for nepali text: a review. Artificial Intelligence Review pp 1–29
28. Nascimento G, Laranjeira C, Braz V, Lacerda A, Nascimento ER (2017) A robust indoor scene recognition method based on sparse representation. CoRR abs/1708.07555
29. Sánchez J, Perronnin F, Mensink T, Verbeek J (2013) Image classification with the fisher vector: theory and practice. Int J Comput Vis 105(3):222–245
30. Li Q, Zhang H, Guo J, Bhanu B, An L (2012) Reference-based scheme combined with k-svd for scene image categorization. IEEE Signal Processing Letters 20(1):67–70
31. Ringnér M (2008) What is principal component analysis? Nature biotechnology 26(3):303–304
32. Bai S, Tang H, An S (2019) Coordinate cnns and lstms to categorize scene images with multi-views and multi-levels of abstraction. Expert Systems with Applications 120:298–309
33. Lin TY, RoyChowdhury A, Maji S (2015) Bilinear cnn models for fine-grained visual recognition. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp 1449–1457
34. Liu S, Tian G, Zhang Y, Duan P (2021) Scene recognition mechanism for service robot adapting various families: A cnn-based approach using multi-type cameras. IEEE Transactions on Multimedia 24:2392–2406

35. Hu J, Guo P (2012) Spatial local binary patterns for scene image classification. In: 2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), IEEE, pp 326–330
36. Banerji S, Sinha A, Liu C (2012) Novel color, shape and texture-based scene image descriptors. In: 2012 IEEE 8th International Conference on Intelligent Computer Communication and Processing, IEEE, pp 245–248
37. Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. Int J Comput Vis 42(3):145–175
38. Zeglazi O, Amine A, Rziza M (2016) Sift descriptors modeling and application in texture image classification. In: Proc. 13th Int. Conf. Comput. Graphics, Imaging and Visualization (CGiV), pp 265–268
39. Wu J, Rehg JM (2011) Centrist: a visual descriptor for scene categorization. IEEE Trans Pattern Anal Mach Intell 33(8):1489–1501
40. Sinha A, Banerji S, Liu C (2014) New color gphog descriptors for object and scene image classification. Machine vision and applications 25(2):361–375
41. Oliva A (2005) Gist of the scene. In: Neurobiology of Attention, Elsevier, pp 251–256
42. Li LJ, Su H, Fei-Fei L, Xing EP (2010) Object bank: A high-level image representation for scene classification & semantic feature sparsification. In: Proc. Adv. Neural Inf. Process. Syst. (NIPS), pp 1378–1386
43. Zhang L, Zhen X, Shao L (2014) Learning object-to-class kernels for scene classification. IEEE Transactions on image processing 23(8):3241–3253
44. Parizi N, Oberlin JG, Felzenszwalb PF (2012) Reconfigurable models for scene recognition. In: Proc. Comput. Vis. Pattern Recognit.(CVPR), pp 2775–2782
45. Juneja M, Vedaldi A, Jawahar C, Zisserman A (2013) Blocks that shout: Distinctive parts for scene classification. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 923–930
46. Lin D, Lu C, Liao R, Jia J (2014) Learning important spatial pooling regions for scene classification. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 3726–3733
47. Quattoni A, Torralba A (2009) Recognizing indoor scenes. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 413–420
48. Zhu J, Li Lj, Fei-Fei L, Xing EP (2010) Large margin learning of upstream scene understanding models. In: Proc. Adv. Neural Inf. Process. Syst. (NIPS), pp 2586–2594
49. ShenghuaGao IH, Liang-TienChia P (2010) Local features are not lonely–Laplacian sparse coding for image classification. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 3555–3561
50. Perronnin F, Sanchez J, Mensink T (2010) Improving the fisher kernel for large-scale image classification. In: Proc. Eur. Conf. Comput. Vis. (ECCV), pp 143–156
51. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. nature 521(7553):436–444
52. Sitaula C, Basnet A, Mainali A, Shahi T (2021) Deep learning-based methods for sentiment analysis on nepali covid-19-related tweets. Computational Intelligence and Neuroscience 2021

53. Sitaula C, Shahi TB (2022) Monkeypox virus detection using pre-trained deep learning-based approaches. Journal of Medical Systems 46(11):1–9
54. Shahi TB, Sitaula C, Neupane A, Guo W (2022) Fruit classification using attention-based mobilenetv2 for industrial applications. Plos one 17(2):e0264,586
55. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 770–778
56. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556 1409.1556
57. Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A (2017) Places: A 10 million image database for scene recognition. IEEE Trans Pattern Anal Mach Intell 40(6):1452–1464
58. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Proc. Adv. Neural Inf. Process. Syst. (NIPS), pp 1097–1105
59. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2014) Going deeper with convolutions. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 1–9, 1409.4842
60. Gong Y, Wang L, Guo R, Lazebnik S (2014) Multi-scale orderless pooling of deep convolutional activation features. In: Proc. Eur. Conf. Comput. Vis. (ECCV), pp 392–407
61. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)
62. Zhou B, Khosla A, Lapedriza A, Torralba A, Oliva A (2016) Places: An image database for deep scene understanding. arXiv preprint arXiv:161002055
63. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: Convolutional architecture for fast feature embedding. In: Proc. 22nd ACM Int. Conf. on Multimedia (ACMM), pp 675–678
64. Kuzborskij I, Maria Carlucci F, Caputo B (2016) When naive bayes nearest neighbors meet convolutional neural networks. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 2100–2109
65. Fornoni M, Caputo B (2014) Scene recognition with naive bayes non-linear learning. In: 2014 22nd International Conference on Pattern Recognition, IEEE, pp 3404–3409
66. Tang P, Wang H, Kwong S (2017) G-ms2f: Googlenet based multi-stage feature fusion of deep cnn for scene recognition. Neurocomputing 225:188–197
67. Zhang C, Zhu G, Huang Q, Tian Q (2017) Image classification by search with explicitly and implicitly semantic representations. Information Sciences 376:125–135
68. Guo Y, Lew MS (2016) Bag of Surrogate Parts: one inherent feature of deep cnns. In: Proc. of the British Machine Vision Conference (BMVC)
69. Gupta S, Dileep AD, Thenkanidiyoor V (2021) Recognition of varying size scene images using semantic analysis of deep activation maps. Machine Vision and Applications 32(2):1–19
70. Zhang C, Liu J, Tian Q, Liang C, Huang Q (2013) Beyond visual features: A weak semantic image representation using exemplar classifiers for classification. Neurocomputing 120:318–324

71. Yang S, Ramanan D (2015) Multi-scale recognition with DAG-CNNs. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp 1215–1223
72. Dixit M, Chen S, Gao D, Rasiwasia N, Vasconcelos N (2015) Scene classification with semantic fisher vectors. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 2974–2983
73. Wang Z, Wang L, Wang Y, Zhang B, Qiao Y (2017) Weakly supervised patchnets: describing and aggregating local patches for scene recognition. IEEE Trans Image Process 26(4):2028–2041
74. Guo S, Huang W, Wang L, Qiao Y (2017) Locally supervised deep hybrid model for scene recognition. IEEE Trans Image Process 26(2):808–820
75. Khan SH, Hayat M, Bennamoun M, Togneri R, Sohel FA (2016) A discriminative representation of convolutional features for indoor scene recognition. IEEE Trans Image Process 25(7):3372–3383
76. Cheng X, Lu J, Feng J, Yuan B, Zhou J (2018) Scene recognition with objectness. Pattern Recognit 74:474–487
77. Lin TYY, RoyChowdhury A, Maji S (????) Bilinear convolutional neural networks for fine-grained visual recognition. IEEE Trans Pattern Anal Mach Intell (6):1309–1322
78. Jiang S, Chen G, Song X, Liu L (2019) Deep patch representations with shared codebook for scene classification. ACM Trans on Multimedia Computing, Communications, and Applications 15(1s):1–17
79. Sorkhi AG, Hassanpour H, Fateh M (2020) A comprehensive system for image scene classification. Multimedia Tools and Applications pp 1–26
80. Chen G, Song X, Zeng H, Jiang S (2020) Scene recognition with prototype-agnostic scene layout. IEEE Trans Image Processing 29:5877–5888
81. Lopez-Cifuentes A, Escudero-Vinolo M, Bescos J, Garcia-Martin A (2020) Semantic-aware scene recognition. Pattern Recognit 102:107,256
82. Zhang B, Wang Q, Lu X, Wang F, Li P (2020) Locality-constrained affine subspace coding for image classification and retrieval. Pattern Recognit 100:107,167
83. Wang D, Mao K (2019) Task-generic semantic convolutional neural network for web text-aided image classification. Neurocomputing 329:103–115
84. Kim Y (2014) Convolutional neural networks for sentence classification. arXiv preprint arXiv:14085882
85. Bosch A, Zisserman A, Muñoz X (2008) Scene classification using a hybrid generative/discriminative approach. IEEE Trans Pattern Anal Mach Intell 30(4):712–727
86. Rasiwasia N, Vasconcelos N (2008) Scene classification with low-dimensional semantic spaces and weak supervision. In: IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 1–6
87. Van Gemert JC, Veenman CJ, Smeulders AW, Geusebroek JM (2009) Visual word ambiguity. IEEE Trans Pattern Anal Mach Intell 32(7):1271–1283
88. Zhang C, Cheng J, Liu J, Pang J, Liang C, Huang Q, Tian Q (2014) Object categorization in sub-semantic space. Neurocomputing 142:248–255
89. Ali N, Zafar B, Riaz F, Dar SH, Ratyal NI, Bajwa KB, Iqbal MK, Sajid M (2018) A hybrid geometric spatial image representation for scene classification. PloS one 13(9):e0203,339
90. Wang D, Mao K (2019) Learning semantic text features for web text-aided image classification. IEEE Trans Multimedia 21(12):2985–2996

91. Griffin G, Holub A, Perona P (2007) Caltech-256 object category dataset
92. Wang J, Wang W, Wang R, Gao W (2016) Csps: An adaptive pooling method for image classification. IEEE Transactions on Multimedia 18(6):1000–1010
93. Sinha A, Banerji S, Liu C (2012) Novel gabor-phog features for object and scene image classification. In: Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), Springer, pp 584–592
94. Silberman N, Fergus R (2011) Indoor scene segmentation using a structured light sensor. In: 2011 IEEE international conference on computer vision workshops (ICCV workshops), IEEE, pp 601–608
95. Ren X, Bo L, Fox D (2012) Rgb-(d) scene labeling: Features and algorithms. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 2759–2766
96. Sun N, Li W, Liu J, Han G, Wu C (2018) Fusing object semantics and deep appearance features for scene recognition. IEEE Transactions on Circuits and Systems for Video Technology 29(6):1715–1728
97. Redmon J, Farhadi A (2017) Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7263–7271
98. Liu S, Tian G (2019) An indoor scene classification method for service robot based on cnn feature. Journal of Robotics 2019
99. Choe S, Seong H, Kim E (2021) Indoor place category recognition for a cleaning robot by fusing a probabilistic approach and deep learning. IEEE Transactions on Cybernetics
100. Fei-Fei L, Perona P (2005) A Bayesian hierarchical model for learning natural scene categories. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. and Pattern Recognit. (CVPR), vol 2, pp 524–531
101. Li LJ, Li FF (2007) What, where and who? classifying events by scene and object recognition. In: Proc. 11th Int. Conf. Comput. Vis. (ICCV), vol 2, p 6
102. Xiao J, Hays J, Ehinger KA, Oliva A, Torralba A (2010) Sun database: Large-scale scene recognition from abbey to zoo. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 3485–3492
103. Niu Z, Zhou Y, Shi K (2010) A hybrid image representation for indoor scene classification. In: 2010 25th International Conference of Image and Vision Computing New Zealand, IEEE, pp 1–7
104. Cho WS, Lam KM (2012) An efficient and effective hybrid pyramid kernel for un-segmented image classification. In: 2012 International Conference on Systems and Informatics (ICSAI2012), IEEE, pp 2153–2158
105. Chen H, Xie K, Wang H, Zhao C (2018) Scene image classification using locality-constrained linear coding based on histogram intersection. Multimedia Tools and Applications 77(3):4081–4092
106. Li Q, Qin Z, Chai L, Zhang H, Guo J, Bhanu B (2013) Representative reference-set and betweenness centrality for scene image categorization. In: 2013 IEEE International Conference on Image Processing, IEEE, pp 3254–3258
107. Sinha A, Banerji S, Liu C (2014) Scene image classification using a wigner-based local binary patterns descriptor. In: 2014 International Joint Conference on Neural Networks (IJCNN), IEEE, pp 1614–1621

108. Zuo Z, Wang G, Shuai B, Zhao L, Yang Q (2015) Exemplar based deep discriminative and shareable feature learning for scene image classification. Pattern Recognition 48(10):3004–3015
109. Liu S, Tian G, Xu Y (2019) A novel scene classification model combining resnet based transfer learning and data augmentation with a filter. Neurocomputing 338:191–206
110. Khan A, Chefranov A, Demirel H (2021) Image scene geometry recognition using low-level features fusion at multi-layer deep cnn. Neurocomputing 440:111–126
111. Liu W, Li Y, Wu Q (2018) An attribute-based high-level image representation for scene classification. IEEE Access 7:4629–4640
112. Qi M, Wang Y (2016) Deep-cssr: Scene classification using category-specific salient region with deep features. In: 2016 IEEE International Conference on Image Processing (ICIP), IEEE, pp 1047–1051
113. Xie GS, Jin XB, Zhang XY, Zang SF, Yang C, Wang Z, Pu J (2018) From class-specific to class-mixture: Cascaded feature representations via restricted boltzmann machine learning. IEEE Access 6:69,393–69,406
114. Bai S (2017) Growing random forest on deep convolutional neural networks for scene categorization. Expert systems with applications 71:279–287
115. Sharma K, Gupta S, Dileep AD, Rameshan R (2018) Scene image classification using reduced virtual feature representation in sparse framework. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 2701–2705
116. Dixit M, Li Y, Vasconcelos N (2019) Semantic fisher scores for task transfer: Using objects to classify scenes. IEEE transactions on pattern analysis and machine intelligence 42(12):3102–3118
117. Lin C, Lee F, Cai J, Chen H, Chen Q (2021) Global and graph encoded local discriminative region representation for scene recognition. Computer Modeling in Engineering & Sciences 128(3):985–1006
118. Wu R, Wang B, Wang W, Yu Y (2015) Harvesting discriminative meta objects with deep cnn features for scene classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1287–1295
119. Wang C, Peng G, De Baets B (2022) Joint global metric learning and local manifold preservation for scene recognition. Information Sciences 610:938–956
120. Aria M, Cuccurullo C (2017) bibliometrix: An r-tool for comprehensive science mapping analysis. Journal of informetrics 11(4):959–975