
# A Systematic Review: Remote Sensed Hyperspectral Image Segmentation and Caption Generation Using Deep Learning Methods

Namdeo Baban Badhe[1(✉)], Vinayak Ashok Bharadi[1], Nupur Giri[2], Sujata Alegavi[3], and Vijaykumar Yele[4]

[1] Department of Information Technology, Finolex Academy of Management and Technology, P-60, P-60/1, Midc, Mirjole Block, Ratnagiri 415639, Maharashtra, India
namdeobadhe1982@gmail.com, vinayak.bharadi@famt.ac.in

[2] Department of Computer Engineering, Vivekanand Education Society's Institute of Technology, Hashu Adwani Memorial Complex, Collector's Colony, Mumbai 400074, Maharashtra, India
nupur.giri@ves.ac.in

[3] Head of the BTech Internet of Things Department, Thakur College Engineering and Technology, Kandivali - (East), Mumbai 400101, India
sujata.alegavi@gmail.com

[4] Electronics and Telecommunication Engineering, Thakur College Engineering and Technology, Kandivali - (East), Mumbai 400101, India
vijaypyele@gmail.com

**Abstract.** Hyperspectral images (HSIs) exhibit a high-dimensional nature, capturing data across numerous wavelengths in the electromagnetic spectrum, often spanning thousands of bands. It has found widespread applications in various real-life scenarios due to its ability to leverage the rich spectral information contained within each pixel. Deep Learning (DL) schemes offer a huge variety of chances to resolve traditional imaging tasks and also for approaching various simulating issues in the spatial-spectral region. This review work provides a systematic review of the relevant existing techniques based on HSI segmentation and image captioning. Initially, other DL methods like, Deep Belief Network (DBN), Convolutional Neural Network (CNN), Autoencoders, Fully Convolutional Neural Network (FCNN), UNet, and Graph Convolutional Network (GCN) are discussed. Secondly, a significant computer vision problem that has recently evolved is image captioning, which tries to automatically produce English explanations of an input image. Therefore, image captioning has garnered growing interest within the realm of remote sensing. This survey summarizes the relevant methods and concentrates on the feature extraction-based methods and attention mechanism-based techniques, which plays a significant role in image caption generation tasks. Finally, it provides the research gaps and its appropriate solution at the end of each survey.

**Keywords:** Hyperspectral image segmentation · remote sensing image captioning · deep learning · feature extraction · attention mechanism


O. Castillo et al. (Eds.): PerSOM 2023, LNICST 517, pp. 39–57, 2024.
https://doi.org/10.1007/978-3-031-66044-3_3

## 1 Introduction

HSIs are typically processed with an extensive number of contiguous narrow spectral wavelengths to analyze terrestrial objects effectively [1]. The extensive hyperspectral bands are harnessed to classify earth classes within HSIs, which are then applied in various distinct applications like verification counterfeit goods and documents, military surveillance, mining, and agriculture etc. [2]. Generally, hyperspectral data is represented as a hypercube $(A * B * C)$. In the context of ground cover analysis, the two dimensions $A$ and $B$ signify spatial information, while the third dimension $(C)$ represents the spectral information. Therefore, various pre-processing functions like atmospheric, radiometric and geometric corrections are used before analysis HSIs further. For the subsequent analysis of the HSI hypercube, it is customary to convert the HSI into a data matrix. [3]. In this transformation, the spectral signature of the vector or pixel is denoted as $a_n$ and it is represented as $a_n = [a_{n1}a_{n2} \cdots \cdots a_{nC}]^T$ in the data matrix where, $n \in [1, R]$, $R$ represents the data matrix, $C$ denotes the number of bands and $R = A * B$.

HSI's segmentation is very difficult because of the complexity involved [4]. To obtain more accurate and easier class identification, researchers have utilized various types of pre-processing methods. Recently, many methods have been introduced for the caption generation in images, which can effectively resolve various computer vision challenges [5]. Based on existing research, similar objects with different spectral values tend to be classified into various classes. Conversely, occasionally, different objects with the same spectral values are grouped together under the same class. It leads misclassification due to concentrating solely over every pixel's spectral vector outcome. To address these complexities, methods based on data extraction and data selection are employed. During the attribute extraction stage, transformations are applied to extract inherent attributes from the HSI. Subsequently, feature selection is employed to choose relevant bands from the HSI data [6]. Multiple feature extraction and selection techniques are utilized to effectively address classification challenges and achieve excellent outcomes. Figure 1 process flow of the review paper.

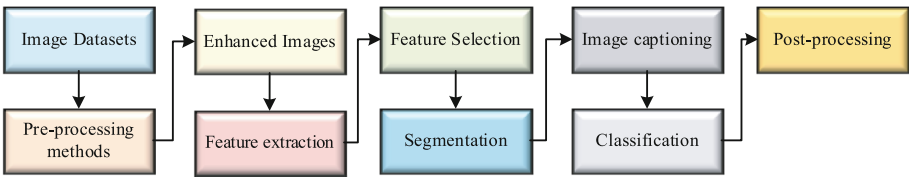


**Fig. 1.** Process flow of the survey

This review paper concentrates on how various existing segmentation strategies have been applied for HSI and different caption generation techniques used for Remote Sensing Images (RSIs) in the existing decade. The performance of the research works performed by the researchers was analyzed by using several evaluation matrices such as, peak signal to-noise ratio, Adjusted Rand Index (ARS), average accuracy, overall accuracy etc.

### 1.1 Contribution of the Review

This paper synthesizes and critically assesses techniques in HSI segmentation and image captioning, especially those utilizing DL. It serves as a comprehensive reference, systematically summarizing past research, identifying gaps, and highlighting emerging trends. This survey focuses on the application of various segmentation and image captioning methods in HSI and remote sensing imagery (RSI) over the past decade. Scholars and professionals' benefit by gaining insights into effective HSI segmentation, facilitating informed decision-making, and guiding future research. Surveys, through their analysis of existing literature, contribute to knowledge advancement, the framing of research questions, and the development of innovative methodologies, ultimately fostering progress in image captioning techniques.

### 1.2 Survey Strategy

This review paper encompasses the process of formulating the survey methodology, conducting a thorough investigation of the chosen method, documenting the obtained outcomes, and exploring the encountered challenges. It also entails describing the sources of information utilized for selection criteria, assessing research articles, evaluating the results, and conducting a quality assessment.

#### 1.2.1 Source of Information

In general, the search process involves seeking conference and journal research papers using resources such as Web of Science, Scopus, academic journals, books, and Google Scholar to retrieve relevant papers. The databases that are utilized in this review work: IEEE Xplore, Springer, ScienceDirect, Google Scholar, Scopus, ACM Digital Library, Taylor & Francis.

This survey is composed of papers published from the years 2019 to 2023, as clearly depicted in Fig. 2.

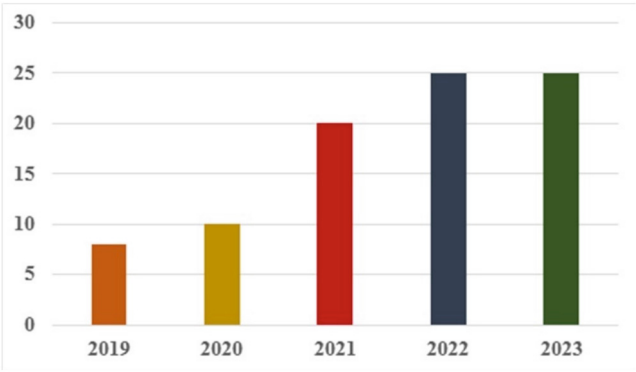


**Fig. 2.** Year wise selection of papers

## 2 Literature Survey

This section provides an explanation of the methods used for image segmentation and caption generation, the commonly applied assessment metrics, and the benchmark datasets utilized in the research.

### 2.1 Existing Segmentation Techniques Used for HSI

The caption generation and classification performance can be simplified enhanced through segmentation methods. Color, a variety of other features, texture and pixel intensity play a vital role in segmentation. There are eight types of segmentation methods are widely considered for the HSI segmentation which are given as follows:

**Thresholding**
This approach is intensity-based, where a specific range of intensities belongs to a similar class, while the rest of the pixels are assigned to another class. A threshold value is employed to differentiate if a pixel represents an object or the background, assigning an intensity value of 0 or 1, respectively. Consider $T$ be the threshold, $h(a, b) = 0$, if $g(a, b) \leq T$ and $h(a, b = 1)$, if $g(a, b) \geq T$ where $g(a, b)$ and $h(a, b)$ is input and output image correspondingly, where $h$ and $g$ are the classes, $a$ and $b$ are the pixels.

**Clustering**
There are two common types of clustering methods are available which are hierarchical and partitional, its groups objects together depend on their similarity or proximity. The hierarchical clustering method builds a tree like cluster's hierarchy. A partition clustering scheme is same as that of k-means clustering, the HSI is segmented into '$k$' clusters and the distance of pixels are calculated based on the seed points of every cluster. The parameters are decided by the shortest distance which are belongs to the same cluster.

**Watershed Segmentation**
The image's gradient is likened to a topographic surface, with bright pixels resembling high points akin to mountaintops or watershed lines, while dark pixels represent low points resembling basins or valleys.

**Morphological Segmentation**
It includes segmenting HSIs by changing its structure and shapes with structural factors. The size of the object boundary is increases, because the morphological dilatation fills holes and gaps in images. The procedure for eliminating the object's outer limit is called as morphological erosion which outcomes in the elimination of minor objects.

**Edge Detection-Based Segmentation**
It works in terms of discontinuity between pixel intensity values and creates binary images. The pixel's first order derivative is compared with a specific threshold value for detecting the image edges. If the second-order derivative exhibits zero crossings, the object boundaries are generated by consolidating the identified edges.

Initially the entire image is treated as a single region in region splitting and merging. If certain similarity constraints are not met, it is then further split into smaller regions. Then, based on homogeneity, regions are merged together.

**DL-Based Segmentation**

Deep learning-based segmentation methods operate on principles inspired by the functioning of the human brain. Within neural networks, multiple hidden nodes are harnessed to capture numerous high-level characteristics that facilitate precise HSI segmentation. Deep learning, CNNs, has made significant strides in various domains of computer vision, including segmentation, detection, and object recognition. These networks take an RGB image as input and execute a series of convolution, pooling, and local normalization functions. The success of deep learning methods in computer vision has not only resonated within the remote sensing community but has also spurred noteworthy advancements in various remote sensing tasks. These tasks encompass very high-resolution satellite image segmentation, hyperspectral image classification, and change detection.

**Existing DL-Based Segmentation Methods Used in HSI**

In 2023, Zhao et al., [7] had proposed Adaptive Superpixel Segmentation (ASS) method to choose the significant samples for Multi-Attention Transformer (MAT). It retains superpixels in uninformative regions while preserving edge information, even in complex regions. This preservation of edge information is leveraged to create favorable local spatial conditions for active learning. In 2023, Fang et al., [8] had introduced an instance segmentation network model for the HSI segmentation. It cannot use both spatial and spectral information effectively. Therefore, the Feature Pyramid Network (FPN) is introduced to integrate multiscale spatial and spectral information during the feature extraction phase. In 2023, Akbari and abkari et al., [9] had proposed an object-based classification method which is a DL model. The weighted Genetic (WG) method was applied to minimize the dimensionality of HSIs. The Expectation Minimization (EM) approach is applied to collect the spatial information. Finally, the segmented image categorized the segmented objects by CNN model.

- **Deep Belief Network (DBN)**

  In 2022, Li et al., [10] had introduced Multi-DBN (MMDBN), to acquire HSI's deep manifold features. To discover the manifold structure present in HSI, a penalty graph and an intrinsic graph are constructed within the manifold layer, assisted by the label information of training samples. Also, it provides the advantages of deep features which enhances the embedding feature's discriminant ability. In 2019, Li et al., [11] had proposed DBN based on multivariate optical sensors and stacked by restricted Boltzmann machine. It classifies the spatial hyperspectral sensor data based on DBN and the feature extraction capability was robust than the other methods.
- **Autoencoders (AEs)**

  In 2020, Nalepa [12] had introduced an unsupervised segmentation to overcome the lack of ground-truth information. Also, proposed 3D convolutional autoencoders

along with clustering strategy which provides end-to-end segmentation results accurately. Tulczyjew et al., [13] had proposed asymmetric AE based on Recurrent Neural Network (RNN), to learn the representation of the compressed unlabeled data and capture both spatial and spectral features in detail. Then, it is integrating with segmentation pipeline which provides high-quality segmentation results.

- **Fully Convolutional Neural Network (FCNN)**

  In 2022, Zaballa et al., [14] had proposed a FCNN for HSI image segmentation. It improves the accuracy of images obtained in real driving environment along with a small-size mosaic snapshot hyperspectral camera. In this proposed system, 3D convolutional methods are used to extract spatial features at various scales. In 2021, Tun et al., [15] had proposed FCNN for HSI classification, after the dimensionality reduction process, it is given to the input layer of CNN. Secondly, the feature extraction was performed which effectively provides the segmentation outcomes with the help of FCNN.

- **UNet based segmentation method**

  In 2022, Li et al., [16] had proposed PSE-UNet system, it is an integration of Principal Component Analysis (PCA) and Squeeze and Excitation (SE) module. Furthermore, it introduced the non-overlapping sliding window scheme, commonly utilized in computer vision, into HSI segmentation. In 2022, Soucy and Sekeh [17] had introduced a Clustering Ensemble U-Net (CEU-Net) for efficient segmentation of HSI. It integrates the extracted spectral information collected from the landscape pixels. Finally, the performance was enhanced by dividing the dataset into same pixels through unsupervised clustering. Figure 3 shows the DL-based image segmentation in HSI

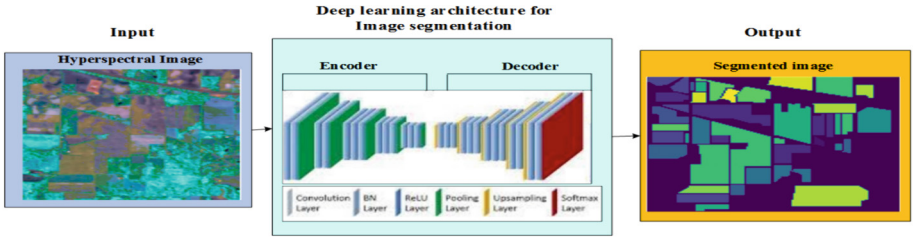


**Fig. 3.** DL-based image segmentation

- **Graph Convolutional Network (GCN)**

  In 2022, Wang and Liang [18] introduced a hybrid approach for HSI classification involving a 3D CNN and scalable GCN. For the feature extraction purpose, a lightweight three-layer CNN was proposed which uses the structural information. Such information was used by the GCN via the similarity matrix. It greatly minimizes the computational complexity of the system. In 2023, Gao and colleagues [19] introduced a fusion network in terms of CNN and GCN, it consists of two stages: a GCN based on superpixel segmentation and CNN with attention mechanism. These two modules extract structural and detailed features from the local region respectively. The obtained

performance was enhanced while both extracted features are got combined. Table 1 provides the existing HSI image segmentation.

**Table 1.** DL-based image segmentation techniques

| Author | Theoretical model | Highlights | Limitations | Datasets and achieved performance |
|---|---|---|---|---|
| Zhao et al., [7] | MAT- adaptive superpixel segmentation | Generate good local spatial conditions for active learning | Required to enhance the generalization performance | • Pavia University (PU): OA is 99.8%<br>• Houston2013: OA is 99.8%<br>• Yellow River Estuary (YRE): OA is 99.8% |
| Fang et al., [8] | Instance segmentation | Effectively differentiate individual instances with same structural information which provides better segmentation outcomes | Failed to segments more complex and diverse practical scenes | • Hyperspectral -Instance Segmentation Dataset HS-ISD: Mask mAP is 62.1 |
| Akbari and abkari et al., [9] | EM model | In the dimensionality reduction approach, no information was deleted and each bands assigned weight among 0 and 1 | Concentrated more on spectral data not on spatial information | • PU: accuracy is 95.8%<br>• DC Mall: accuracy is 96.9%<br>• Indian Pine (IP): accuracy is 93.2% |
| Zaballa et al., [14] | FCN segmentation | Major objective is to learn, what extent the spatial features are codified by the convolutional filters | Required proper knowledge about the data transfer to enhance the throughput further | • HSI-Drive v1.1: accuracy is 95.37% precision is 95.55% IoU is 91.31% |

(*continued*)

**Table 1.** (*continued*)

| Author | Theoretical model | Highlights | Limitations | Datasets and achieved performance |
|---|---|---|---|---|
| Nalepa [12] | Unsupervised segmentation | Provides consistent and great quality segmentation results without used any labels | Computational cost was maximum due to the dimension of features | • IP:<br>Normalized Mutual Information (NMI) is 0.431<br>ARS is 0.231<br>• PU:<br>NMI is 0.553<br>ARS is 0.339<br>• SA:<br>NMI is 0.714<br>ARS is 0.533<br>• Mullewa dataset<br>NMI is 8.33<br>• ARS is 8.00 |
| Li et al., [10] | MMDBN | The extracted abstract features signify the deep information | Spatial features are not considered | • IP: accuracy is 81.50%<br>• SA: accuracy is 91.79%<br>• Botswana HSI: accuracy is 94.05% |
| Gao et al., [19] | Fused CNN with GCN | This method better represents the features of the nodes | The variability of neighboring nodes was not considered | • IP: accuracy is 98.78%<br>• PU: accuracy is 98.99%<br>• SA: accuracy is 98.69% |

Table 2 enumerate the research gaps encountered in prior research and provide corresponding solutions to address these challenges.

**Table 2.** Research gap and solution on various segmentation algorithms

| Research gap | Solution |
|---|---|
| Because of the varying circumstances affecting HSI data, it is predominant to increase the testing accuracy | By trained well DL systems capable of identify unseen testing images accurately, which is referred as Generalization |
| If hyperspectral image segmentation is purely based on spectral information, leads to high proportion of false positives | Combine spectral information with other imaging modalities like LiDAR or multispectral data to enhance segmentation accuracy and reduce false positives by leveraging complementary information |

## 2.2 Caption Generation for RSIs

According to the content observed in an image, the caption automatically generating natural language descriptions It is a significant part of scene understanding that integrates the knowledge of natural language processing and computer vision. The application of image caption is significant and extensive, for example human-computer interaction's realization. In RSIs, the image captioning is a complex issue, where the work is to generate a description of the provided RSI. In this review, few recent images captioning research works have discussed by applying feature extraction-based methods and attention mechanism. Hence, a DL technique is employed to further augment the caption generation process.

**DL Based RSI Caption Generation**

Now-a-days, DL based caption generation approaches, the system is built in terms of encoder-decoder network design. In the encoding phase, the input image's high-level internal representations are extracted by using the deep CNNs. In the decoding phase, the internal representations are decoded into sentence descriptions using a trained RNN.

### 2.2.1 Existing Feature Extraction-Based Methods for Caption Generation

In 2022, Wang et al., [20] had proposed multiscale multi-interaction feature extraction module for the efficient caption generation. It is a two-stage process, at first finetune the neural network backbone over RSIs. Here, grasses and low plants identification was difficult. Secondly, a multi-interaction feature representation model was proposed to compute the similarity score between features. In 2023, Zhao et al., [21] had introduced Dual Feature Enhancement Network ("DFEN") based on codec to enhance object information at both the text and image levels. Furthermore, with the assistance of the image enhancement module, huge discriminative context features are achieved. It uses visual features of images to handle the text-enhancement system, resulting in text-guided features that accurately focus on the ground object's information.

### 2.2.2 Existing Attention Mechanism-Based Methods for Caption Generation

Another one RSIC technique had presented in 2019, Zhang et al., [22] with Label-Attention Mechanism (LAM). The presented approach generates high quality sentences with label information from high resolution images. The generated sentences were fully based on the image features thus, the text has pure and useful information. Figure 4 illustrates the DL-based caption generation structure. A multi-level attention model had integrated in 2020, Li et al., [23] to achieve accurate caption generation from RSI. There are three different types of attention mechanisms were integrated to achieve captions based on different area and different vision of images.

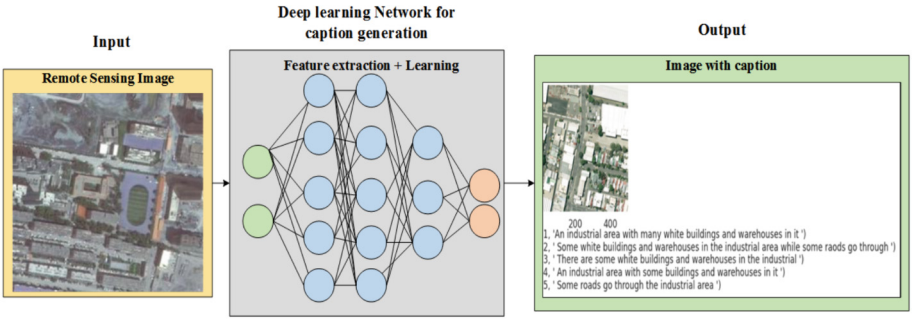


**Fig. 4.** RSI caption generation based on DL

In 2022, Gajbhiye and Nandedkar [24] had proposed MEmory-guided Transformer (SCAMET) framework based on Spatial-Channel Attention to generate accurate captions. The DL framework of CNN was integrated with this transformer. In 2023, Zhang et al., [25] had introduced a caption generation technique based on the visual content and ROI in the RSI. The stair attention mechanism, which facilitates interaction among multiple sources, propels the entire process. The caption generation search the image in three different regions which are core, surrounding and other regions. Then the quality of information was improved by CIDEr-based reward reinforcement learning.

**Convolutional Neural Network Based Caption Generation**

In 2022, Wang et al., [26] had introduced a caption generation technique from RSI with Global–Local Captioning Model (GLCM) to achieve accurate feature representation. The generated words from all the visual texture of images are related to each other and separate words based on the relevant visual features.

**Graph Convolutional Module for Caption Generation**

In 2019, Yuan et al., [27] had proposed multi-label attribute graph and multi-level attention model convolutional. Initially, a multi-level attention system can focus not only on specific spatial features but also on features at particular scales. Secondly, multi-label attribute graph convolutional method was used to learn more efficient features for caption generation.

**Transformer Based Image Caption**

Ren et al., [28] had introduced mask guided transformer network to enhance the performance and caption generation. A token, which is a hybrid of the encoder, represents the scene's topic and assumes a significant role in the decoder module. In order to enhance the caption's diversity, mask cross entropy concept was applied. In 2023, Chang et al., [29] had introduced Changes-to-Captions (Chg2Cap) to generate accurate captions. It undergoes three stages: Siaseme CNN based feature extractor, an attentive decoder and transformer-based caption generator.

**LSTM Based Caption Generation**

In 2023, Xie et al., [30] introduced a Bidirectional LSTM and Attention Mechanism (Bi-LS-AttM) to improve the generation of image captions. Further, improve the accuracy and temporal efficiency, this system uses CNN along with Fast Region-based CNN (FRCNN) which is used for feature extraction purpose. Table 3 list out the existing caption generation techniques. Table 5 provides the achieved results of the existing methods based on three datasets such as, UCM, Sydney and RSICD.

Table 4 enumerate the challenges encountered in the existing research and outline the corresponding solutions to address these gaps.

**Table 3.** DL based remote sensing caption generation

| Author | Techniques | Highlights | Input | Comment | Dataset |
|---|---|---|---|---|---|
| Wang et al., [20] | Multiscale multi-interaction feature extraction | Different scale images captions are handled efficiently. |  | Some driving cars on two curved freeways | UCM-Captions, RSICD and Sydney-Captions |
| Zhao et al., [21] | DFEN | High-resolution images captions were automatically generated by UAV inspections |  | Several green trees and meadows are in two sides of a green river | UCM-Captions, RSICD and Sydney-Captions |
| Xie et al., [30] | Fast RCNN | The model applicability was enhanced by integrating multitask learning methods. |  | Certain roads pass through industrial areas replete with numerous buildings and warehouses. | Flickr30K and MSCOCO |
| Wang et al., [26] | GLCM | The better feature representation was done with up-bottom strategy |  | Some storage tanks are near a piece of bareland | UCM-Captions, RSICD and Sydney-Captions |
| Li et al., [23] | Multi-Level Attention Model | The visual feature extraction was effectively increasing the image description. |  | Green trees and several buildings surround a church. | UCM-Captions, RSICD and Sydney-Captions |
| Gajbhiye and Nandedkar [24] | SCAMET | Provides high-level semantic information. |  | There are some buildings with grey roof and parking lot | UCM-Captions, RSICD and Sydney-Captions |

**Table 4.** Research gap and solution on various segmentation algorithms

| Research Gap | Solution |
|---|---|
| RSIs can be complex and contain multiple regions of interest. Generating coherent and informative captions that describe all relevant details becomes challenging, especially for long captions | Incorporate attention mechanisms into the captioning models to concentrate on different image regions during caption generation, thereby enhancing the model's ability to provide more precise descriptions of specific areas |

### 2.3 Implementation Using Datasets

In this section, descriptions of the datasets used are provided, along with the introduction of multiple evaluation metrics. Subsequently, the training process is explained comprehensively, offering insights into the methodology. Finally, the section presents the experimental results and conducts in-depth analyses.

**UCM Dataset**
The UCM dataset is a renowned repository of remote sensing images [21], primarily tailored for land use classification. It encompasses high-resolution satellite images depicting a diverse range of urban and rural landscapes. Thoroughly annotated with precise ground truth data, these images serve as valuable resources for training and assessing machine learning algorithms, particularly for tasks like classifying land cover and analyzing land use. The dataset encompasses nearly 21 distinct categories, each containing a substantial number of images, all standardized to a resolution of $256 \times 256$ pixels. Additionally, each image within the dataset is accompanied by five descriptive sentences.

**Sydney-Caption Dataset**
The Sydney Captions Dataset is a curated compilation of concise and descriptive captions paired with images of Sydney, Australia. Developed to enhance content with meaningful captions, this dataset proves invaluable for automating image captioning tasks. It provides textual descriptions for various scenes, landmarks, and aspects of Sydney, making it a valuable resource for AI and language models. This dataset comprises seven categories and includes a total of 613 images [21], each with a size of $500 \times 500$ pixels. Additionally, every image within the dataset is paired with five distinct natural language descriptions.

**RSICD (Remote Sensing Image Captioning Dataset)**
The RSICD is a dedicated dataset for remote sensing images. It features a diverse array of high-resolution images captured by satellites, accompanied by detailed and informative captions. RSICD plays a pivotal role in training and evaluating image captioning models specifically designed for remote sensing applications. This dataset enables advanced comprehension and interpretation of remote sensing data. RSICD is the largest among the three datasets [21], encompassing a total of 10,921 images distributed across 30

categories. The images have a size of 224 × 224 pixels, and each image is annotated with artificial annotations consisting of five sentences.

**Evaluation Metrics**
In the realm of machine translation and natural language processing, two prominent evaluation metrics stand out: BLEU and ROUGE. These metrics play a pivotal role in assessing the quality of generated text, especially when comparing machine-generated translations to their reference, which is typically created by humans.

**BLEU (Bilingual Evaluation Understudy)**
It is a metric created to measure the similarity between translations generated by machines and one or more reference translations. It offers a comprehensive evaluation by considering various subsets. BLEU calculates an overall score by amalgamating the precision of n-grams from these subsets. Higher BLEU scores signify superior translation quality. Nonetheless, BLEU has its limitations, such as its inability to capture semantic nuances. The BLEU subsets include:

- **BLEU-1:** Evaluates the precision of unigrams (single words) in the text generated by the machine in comparison to the reference text
- **BLEU-2:** Measures the precision of bigrams (pairs of consecutive words) in the text generated by the machine in comparison to the reference text
- **BLEU-3:** Assesses the precision of trigrams (triplets of consecutive words) in the text generated by the machine in comparison to the reference text
- **BLEU-4:** Gauges the precision of four-grams (quartets of consecutive words) in the text generated by the machine in comparison to the reference text.

**ROUGE (Recall-Oriented Understudy for Gisting Evaluation)**
It serves as a unit to gauge the fidelity of text summaries or machine-generated content by juxtaposing it with one or more reference summaries. It offers a multifaceted evaluation by employing diverse measures, including ROUGE-N (precision and recall of n-grams), ROUGE-W (weighted longest common subsequence), ROUGE-L (longest common subsequence), and among others. ROUGE emphasizes both recall (the degree to which the reference summary is captured) and precision (the relevance of the generated content), thus providing a more thorough assessment of text quality.

**Table 5.** Accomplished performance measures of existing image captioning techniques

| UCM dataset | | | | | | |
|---|---|---|---|---|---|---|
| DL | Existing methods | BLEU 1 | BLEU 2 | BLEU 3 | BLEU 4 | ROUGE |
| Feature extraction-based | Wang et al., [20] | 0.843 | 0.775 | 0.711 | 0.651 | 0.785 |
| | Zhao et al., [21] | 0.851 | 0.784 | 0.728 | 0.677 | 0.805 |
| Attention based | Zhang et al., [22] | 0.857 | 0.812 | 0.775 | 0.743 | 0.826 |
| | Li et al., [23] | 0.80358 | 0.73616 | 0.68453 | 0.63829 | 0.76923 |
| | Gajbhiye and Nandedkar [24] | 0.8460 | 0.7772 | 0.7262 | 0.6812 | 0.8166 |
| | Zhang et al., [25] | 0.8727 | 0.8096 | 0.7551 | 0.7039 | 0.8258 |
| CNN | Wang et al., [26] | 0.8182 | 0.7540 | 0.6986 | 0.6468 | 0.7524 |
| Graph Convolutional | Yuan et al., [27] | 0.8330 | 0.7712 | 0.7154 | 0.6623 | 0.7763 |
| Transformer | Ren et al., [28] | 89.36 | 84.82 | 80.57 | 76.50 | 85.86 |
| Sydney-caption | | | | | | |
| Feature extraction-based | Wang et al., [20] | 0.842 | 0.757 | 0.672 | 0.601 | 0.733 |
| | Zhao et al., [21] | 0.798 | 0.697 | 0.614 | 0.542 | 0.723 |
| Attention based | Zhang et al., [22] | 0.7365 | 0.6440 | 0.5835 | 0.5348 | 0.6827 |
| | Li et al., [23] | 0.72743 | 0.63837 | 0.56260 | 0.50244 | 0.71541 |
| | Gajbhiye and Nandedkar [24] | 0.8072 | 0.7136 | 0.6431 | 0.5846 | 0.7258 |
| | Zhang et al., [25] | 0.7643 | 0.6919 | 0.6283 | 0.5725 | 0.7172 |
| CNN | Wang et al., [26] | 0.8041 | 0.7305 | 0.6745 | 0.6259 | 0.6965 |
| Graph Convolutional | Yuan et al., [27] | 0.8233 | 0.7548 | 0.6587 | 0.6003 | 0.7237 |
| Transformer | Ren et al., [28] | 83.38 | 75.72 | 67.72 | 59.8 | 76.60 |
| RSICD | | | | | | |
| Feature extraction-based | Wang et al., [20] | 0.793 | 0.681 | 0.577 | 0.498 | 0.682 |
| | Zhao et al., [21] | 0.766 | 0.636 | 0.538 | 0.463 | 0.685 |

(*continued*)

**Table 5.** (*continued*)

| RSICD | | | | | | |
|---|---|---|---|---|---|---|
| Attention based | Zhang et al., [22] | 0.6756 | 0.5549 | 0.4714 | 0.4077 | 0.5848 |
| | Li et al., [23] | 0.75799 | 0.60242 | 0.49857 | 0.42243 | 0.67660 |
| | Gajbhiye and Nandedkar [24] | 0.7681 | 0.6309 | 0.5352 | 0.4611 | 0.6979 |
| | Zhang et al., [25] | 0.7836 | 0.6679 | 0.5774 | 0.5042 | 0.6730 |
| CNN | Wang et al., [26] | 0.7767 | 0.6492 | 0.5642 | 0.4937 | 0.6779 |
| Graph Convolutional | Yuan et al., [27] | 0.7597 | 0.6421 | 0.5517 | 0.4623 | 0.6563 |
| Transformer | Ren et al., [28] | 80.42 | 80.42 | 61.36 | 54.14 | 70.58 |

## 3 Challenges and Future Scope

The various challenges and recommended future scope while handling segmentation and image captioning are given as follow:

***HSI Segmentation***

- HSI represents an evolving field of study. The intricate nature of hyperspectral data poses challenges for conventional ML techniques when it comes to segmentation. More recently, DL has emerged as a potent tool, delivering state-of-the-art results in various applications. For the segmentation and image captioning, encouraging outcomes have been found in RS by applying DL schemes. Therefore, using DL approaches in the advanced application domains is an exciting technology to survey.
- In the DL systems, scarcity of publicly accessible datasets and their limited size leads to poor performance or overfitting issues. It is a promising avenue for future research aimed at introducing innovative methods to support various applications that demand a large-scale hyperspectral database.
- Spaceborne sensors may capture data that leads to a mixed pixel effect, introducing complexity and challenges in segmentation work. Consequently, there is a demand for the development of an algorithm capable of automatically detecting mixed pixels.
- The vast volume of data captured from hundreds of spectral bands is stored in data cubes. Many of these bands exhibit significant correlation, leading to redundant data. It requires automatic protocols to separate the redundant bands which can improve accuracy. To overcome the above-mentioned challenges, design an innovative Optimized Segmentation Network (OptSegNet) based RSI segmentation is introduced, which is a dual-path Resnet_50 with UNet and a convolutional network hybridization. The proposed method, illustrated in Fig. 5, comprises four stages: (i) preprocessing, (ii) feature extraction, (iii) segmentation, and (iv) post-processing.

- Initially, the input RSIs are sourced from hyperspectral datasets, including IP, PU, and SD. Secondly, these images undergo preprocessing through guided box filtering before being fed into OSegNet for feature extraction and segmentation.
- To augment the performance of OptSegNet, EMGO is employed. Subsequently, the segmented image is processed further using Pairwise Neural Conditional Random Field (PNCRF) for enhanced segmentation accuracy.
- Finally, the introduced model is compared with several previous models to demonstrate the effectiveness of the feature extraction-based segmentation algorithm.

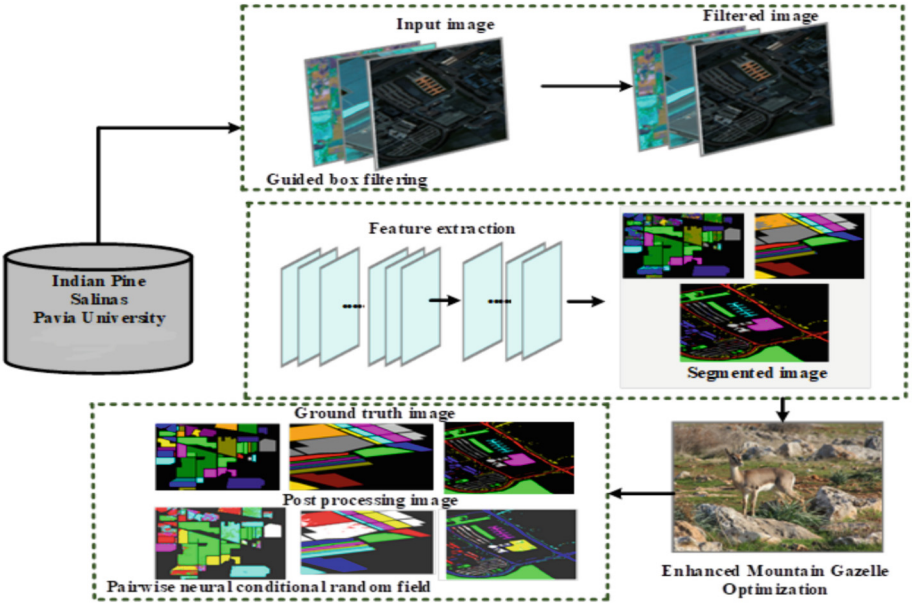


**Fig. 5.** Proposed HSI Segmentation based on OptSegNet Architecture

### *RSI Caption Generation*

- Dimensionality reduction is a significant factor required to be addressed. Novel approaches are need to be designed to eliminate the irrelevant features which would be beneficial for analysis.
- The system is designed to possess the capability to generate description sentences respective to multiple main objects instead of describing a single target object.
- Common image description model should be designed for handling multiple languages and also optimize the acceleration of testing, training and generating sentences to enhance the accuracy performance. To overcome the above-mentioned challenges, proposed a Deep Attention applied DenseNet with visual switch added (DADN-BiLSTM) for captioning shown in Fig. 6.

- Initially, RSIs for input are sourced from various datasets and preprocessed using Improved Gaussian Rolling Guidance Filter (IGRGF). Then the captions are preprocessed using some methods.
- These preprocessed images are given to the Double Attention-based DenseNet ($A^2$DenseNet) to extract different scale image features to give whole notation of image which played as encoder in the research.
- Then, an Adaptive K-Dimensional Tree (AKD-Tree) based Euclidean clustering is utilized to segment the image depends on extracted features.
- Then the segmented image and preprocessed captions are given to BiLSTM, that applied as the decoder to improve the use of context information. Finally, the corresponding captions with image is obtained in the output.

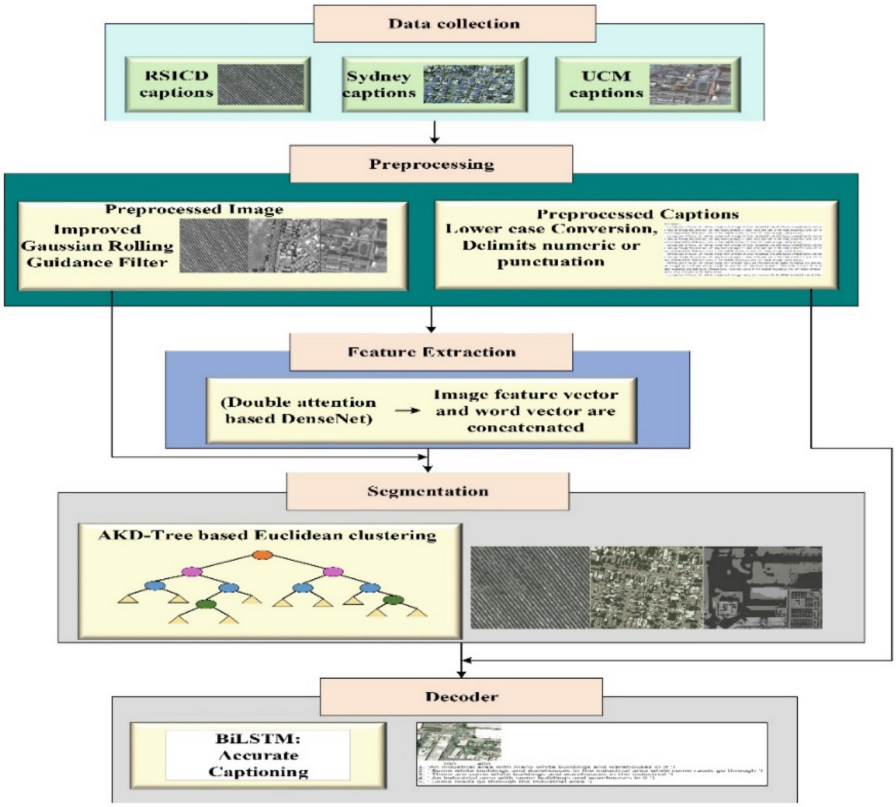


**Fig. 6.** Work flow of proposed DADN- BiLSTM based RSI captioning

## 4 Conclusion

This survey reviews the various methods which have been widely distinguished according to several techniques such as DL based, feature extraction and attention mechanism-based methods. Few researchers have gathered their real time HSI images of vegetables and fruits. Various researchers applied the standard databases like IP, PU, SA, UCM-caption, Sydney-caption, and RSICD. These benchmark datasets handle multiple classes of land cover such as shelter, roads and vegetations. The accuracy performance achieved by various approaches shows that DL-based approaches outperform the non-DL-based approaches. In Unet based segmentation, choose only the most relevant features and minimizing the number of features can prevent the model from fitting noise in the data which is performed in the proposed OptsegNet's max pooling layer. In attention mechanism-based image captioning, feature extraction concentrates on multiscale features of the objects in the images. Therefore, caption is generated for the multiple objects instead of single objects. From the experimental results, it is clear that the Unet and attention mechanism achieves greatest enhancement among all techniques. This experimental outcome may provide few guidelines for the future learning on this topic.

## References

1. Uddin, M.P., Mamun, M.A., Hossain, M.A.: PCA-based feature reduction for hyperspectral remote sensing image classification. IETE Tech. Rev. **38**, 377–396 (2020)
2. Uddin, M.P., Mamun, M.A., Hossain, M.A.: Effective feature extraction through segmentation-based folded-PCA for hyperspectral image classification. Int. J. Remote Sens. **40**, 7190–7220 (2019)
3. Afjal, M.I., Mondal, M.N., Mamun, M.A.: Segmentation-based linear discriminant analysis with information theoretic feature selection for hyperspectral image classification. Int. J. Remote Sens. **44**, 3412–3455 (2023)
4. Kumar, G., Kumar, A., Singhal, M., Singh, K.U., Kumar, L., Singh, T.: Revolutionizing plant disease management through image processing technology. In: 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES) (2023)
5. Chen, Z., Wang, J., Ma, A., Zhong, Y.: Typeformer: multiscale transformer with type controller for remote sensing image caption. IEEE Geosci. Remote Sens. Lett. **19**, 1–5 (2022)
6. Islam, M.R., Ahmed, B., Hossain, M.A., Uddin, M.P.: Mutual information-driven feature reduction for hyperspectral image classification. Sensors **23**, 657 (2023)
7. Zhao, C., et al.: Hyperspectral image classification with multi-attention transformer and adaptive superpixel segmentation-based active learning. IEEE Trans. Image Process. **32**, 3606–3621 (2023)
8. Fang, L., Jiang, Y., Yan, Y., Yue, J., Deng, Y.: Hyperspectral image instance segmentation using spectral–spatial feature pyramid network. IEEE Trans. Geosci. Remote Sens. **61**, 1–13 (2023)
9. Akbari, D., Akbari, V.: Object-based classification of hyperspectral images based on weighted genetic algorithm and deep learning model. Appl. Geomatics **15**, 227–238 (2023)
10. Li, Z., Huang, H., Zhang, Z., Shi, G.: Manifold-based multi-deep belief network for feature extraction of hyperspectral image. Remote Sens. **14**, 1484 (2022)
11. Li, C., Wang, Y., Zhang, X., Gao, H., Yang, Y., Wang, J.: Deep belief network for spectral–spatial classification of hyperspectral remote sensor data. Sensors **19**, 204 (2019)

12. Nalepa, J., Myller, M., Imai, Y., Honda, K.-I., Takeda, T., Antoniak, M.: Unsupervised segmentation of hyperspectral images using 3-D convolutional autoencoders. IEEE Geosci. Remote Sens. Lett. **17**, 1948–1952 (2020)
13. Tulczyjew, L., Kawulok, M., Nalepa, J.: Unsupervised feature learning using recurrent neural nets for segmenting hyperspectral images. IEEE Geosci. Remote Sens. Lett. **18**, 2142–2146 (2021)
14. Gutiérrez-Zaballa, J., Basterretxea, K., Javier Echanobe, M., Martínez, V., del Campo, I.: Exploring fully convolutional networks for the segmentation of hyperspectral imaging applied to advanced driver assistance systems. In: Desnos, K., Pertuz, S. (eds.) DASIP 2022. LNCS, vol. 13425, pp. 136–148. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-12748-9_11
15. Tun, N.L., Gavrilov, A., Tun, N.M., Trieu, D.M., Aung, H.: Hyperspectral remote sensing images classification using fully convolutional neural network. In: 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus) (2021)
16. Li, J., Wang, H., Zhang, A., Liu, Y.: Semantic segmentation of hyperspectral remote sensing images based on PSE-UNET model. Sensors **22**, 9678 (2022)
17. Soucy, N., Sekeh, S.Y.: CEU-Net: ensemble semantic segmentation of hyperspectral images using clustering. J. Big Data **10**, 43 (2023)
18. Wang, X., Liang, Z.: Hybrid network model based on 3D convolutional neural network and scalable graph convolutional network for hyperspectral image classification. IET Image Process. **17**, 256–273 (2022)
19. Gao, L., Xiao, S., Hu, C., Yan, Y.: Hyperspectral image classification based on fusion of convolutional neural network and graph network. Appl. Sci. **13**, 7143 (2023)
20. Wang, Y., Zhang, W., Zhang, Z., Gao, X., Sun, X.: Multiscale multiinteraction network for remote sensing image captioning. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **15**, 2154–2165 (2022)
21. Zhao, W., Yang, W., Chen, D., Wei, F.: DFEN: dual feature enhancement network for remote sensing image caption. Electronics **12**, 1547 (2023)
22. Zhang, Z., Diao, W., Zhang, W., Yan, M., Gao, X., Sun, X.: LAM: remote sensing image captioning with label-attention mechanism. Remote Sensing. **11**, 2349 (2019)
23. Li, Y., Fang, S., Jiao, L., Liu, R., Shang, R.: A multi-level attention model for remote sensing image captions. Remote Sens. **12**, 939 (2020)
24. Gajbhiye, G.O., Nandedkar, A.V.: Generating the captions for Remote Sensing Images: a spatial-channel attention based memory-guided transformer approach. Eng. Appl. Artif. Intell. **114**, 105076 (2022)
25. Zhang, X., et al.: Multi-source interactive stair attention for remote sensing image captioning. Remote Sens. **15**, 579 (2023)
26. Wang, Q., Huang, W., Zhang, X., Li, X.: GLCM: global–local captioning model for remote sensing image captioning. IEEE Trans. Cybern. **53**(11), 6910–6922 (2022)
27. Yuan, Z., Li, X., Wang, Q.: Exploring multi-level attention and semantic relationship for remote sensing image captioning. IEEE Access **8**, 2608–2620 (2020)
28. Ren, Z., Gou, S., Guo, Z., Mao, S., Li, R.: A mask-guided transformer network with topic token for remote sensing image captioning. Remote Sens. **14**, 2939 (2022)
29. Chang, S., Ghamisi, P.: Changes to captions: an attentive network for remote sensing change captioning. arXiv preprint arXiv:2304.01091 (2023)
30. Xie, T., Ding, W., Zhang, J., Wan, X., Wang, J.: Bi-LS-AttM: a bidirectional LSTM and attention mechanism model for improving image captioning. Appl. Sci. **13**, 7916 (2023)