Redesigning U-Net with dense connection and attention module for satellite based cloud detection

Aarti Kumthekar[1], Gudheti Ramachandra Reddy[1,2]
[1]Department of Electronics Engineering, Vellore Institute of Technology, Vellore, India
[2]Piscataway, New Jersey, United States of America

Article Info

Article history:

Received Oct 15, 2021
Revised Jan 22, 2022
Accepted Feb 6, 2022

Keywords:

Attention module
Cloud detection
Deep learning
Dense connection

ABSTRACT

In this paper, we present an upgraded U-Net technique for satellite-based cloud detection, with additional features, such as, more relevant spatial information, improvement in gradient propagation, feature reuse and controlling the network parameters using growth rate by adding dense connections. Furthermore, incorporation of attention module helps to learn strong inter-spatial and inter-channel relationships of feature maps by adding a few trainable parameters to the network. The two attention blocks namely position attention module (PAM) and channel attention module (CAM) focus on important parts of the image by neglecting the redundant information. The experimental results prove that the put forward technique with dense and attention modules could detect cloud with an accuracy of 95.69%.

Corresponding Author:

Gudheti Ramachandra Reddy
School of Electronics Engineering, Vellore Institute of Technology
Vellore, India
Email: grreddy@vit.ac.in

1. INTRODUCTION

Satellite remote sensing (RS) is the science of acquisition of knowledge regarding areas from satellite by capturing the energy emitted or reflected from earth's surface. Remote sensing images contain diverse information which is useful in understanding and exploring environmental changes, resource management, natural calamities management, and object recognition. There are various types of satellites for observing the earth surface and to capture various information with respect to sensor payload, orbit and different resolution (spectral, temporal or spatial). Based on necessity and applications, orbits have different revisit frequency (low and high), resolution (low and high) and wide swaths. One of the longest-running satellite is Landsat satellites which enables users to utilize the earth's observation information. Every 16 days, Landsat 8 satellite sensors capture images with different bands such as multispectral bands from operational land imager (OLI) and two bands from thermal infrared sensors (TIRS). However, the interpretation, analysis, and utilization of remote sensing images suffer greatly due to cloud coverage. The cloudy image causes difficulty in acquiring complete information of earth's land surface area. The requirement is to obtain remote sensing images without any cloud coverage, but it is merely impossible to capture cloud free images. With the aim to improve the interpreability and analysis of remote sensing images, indentification of the cloud is a vital task. Accurate, automatic and reliable image segmentation is a vital pre-processing step for interpretation and investigation of remote sensing images in different applications [1]–[7].

In recent times, different algorithms have been introduced for cloud detection which are broadly classified as traditional and algorithm based. The traditional algorithms are simple threshold based methods [8]–[11] or statistical based methods such as histogram [12], clustering [13] and textures [14]–[17].

The cloud, cloud shadow and detection of snow can be classified with some of the effective traditional algorithms such as automated cloud cover assessment (ACCA) [18], function of mask (F-Mask) [19] and haze optimized transformation (HOT) [20]. The cloud detection is a more challenging task and these methods are limited in utility and have to overcome the drawbacks; thus, the machine learning techniques have evolved with better efficiency. Machine learning techniques such as support vector machine, random forest, discriminant analysis, Markov fields, Naïve Bayes, nearest neighbour and others have promising performance. As the cloud detection methods based on machine learning use hand-crafted features that are tedious, error prone and time consuming, the automatic feature extraction based deep learning techniques have evolved for accurate cloud detection.

In last decade, a deep learning (DL) has made tremendous breakthroughs in many remote sensing applications. convolutional neural network (CNN) approach that uses patch-to-pixel or encoder-decoder segmentation architectures have proven successful due to the availability of the increasing computational power and their inherent ability to perceive spatial information. The algorithm which is combination of machine learning techniques and superpixel algorithm [21] is further modified with deep learning. Instead of using machine learning algorithm for classification, CNN is utilized to classify the cloud in the image [22]. A multilevel cloud detection was designed by Xie *et al.* [23] with improved superpixel segmentation and two branch convolutional neural network to improve the accuracy. Further, [24] used adaptive simple linear iterative clustering with multiple convolutional neural networks which extract multiscale features and multilevel clouds to detect thin, thick and non-cloud areas. However, the initial superpixel segmentation should be correctly done in order to get good results and also the performance of algorithm depends on size of input image. All these techniques are based on superpixel segmentation instead of pixel-wise segmentation.

In last few years, CNN is widely used in semantic segmentation which is the task of designating a class label to each pixel in the image. Instead of fully connected layer, the network utilizes stack of convolution layers to maintain same input and output size. The pixel-wise segmentation using encoder decoder structure helps increasing the model accuracy and also reduces the number of parameters. Fully convolutional network (FCN) [25], U-Net [26], SegNet [27], and deeplab are widely utilised methodologies for semenatic based image segmentation. Generally, the structure is made up of both convolution and transposed convolution which helps to learn the semantic transformation between input and output image. Their performance is outstanding in many applications because of their automated feature extraction. Motivated by these networks, researchers are utilizing these techniques in the remote sensing applications. Fully convolutional network is implemented to distinguish the cloud and snow from the multispectral images, by integrating the spatial information and semantic information with a multiscale prediction module [28]. Similarly using FCN [29] presence of cloud is detected in the Landsat 8 four spectral band red, green, blue and near infra-red (RGBNIR) image. Modification to SegNet and GoogleNet techniques helped to emerging P_SegNet and NP_SegNet based algorithms [30]. On spatial procedures for automated removal of cloud and shadow (SPARCS) landsat dataset, [31] developed the simple lightweight based U-Net technique combined with Legall 5/3 wavelet transform. The cloud and shadow of Landsat 8 and GF-1 satellite dataset was detected by authors in [32] with encoder decoder structure and feature pyramid module and boundary refinement. RS-Net [33] was developed based on U-Net architecture to detect the cloud in Landsat 8 dataset. Although satisfactory results are obtained using above mentioned methods, still significant research is progressing using U-Net as the backbone architecture. In the proposed work, U-Net is utilized as basic model, further it is modified to get more robust results. The main novelty of our proposed algorithm is as:

i) We implemeted an architecture which utilizes dense connection for maximum flow of information between the layers in the encoder-decoder structure of U-Net. It combines both low-level high-resolution features and high-level semantic strong features to segment the image more accurately.

ii) We also integrate the attention module into the network to focus on more powerful inter spatial and inter spectral based representative features. By neglecting irrelevant details, it improves the discriminate ability between non-cloud and cloudy pixels.

iii) The proposed algorithm is evaluated on benchmark dataset for cloud understanding without using pre-trained parameters or post-processing.

The paper is organized: section 2 presents the detailed framework of dense module and attention module. In section 3, the description of the information of dataset, experimental results and cloud detection performance are discussed. The final conclusions of this paper are presented in section 4.

2. RESEARCH METHOD

Ronneberger *et al.* [26] developed U-Net architecture which is based on CNN for bio-medical image segmentation which is combination of convolution layer and maxpooling layer. As the convolutional layer depth increases, the semantic information of the image increases which extracts more relevant and accurate features. The U-Net architecture is an encoder and decoder based structure with contractive and expansive parts that combines higher-level detail information and lower-level semantic information from the satellite image. Generally, U-Net architecture has four down-sampling and up-sampling blocks with convolution along with batch normalisation and ReLU layer with lesser number of parameters. The U-Net architecture can be improved by making the network deeper and wider, which allows the network to adapt more meaningful features that retain lower-level details along with extraction of high-level semantic feature information which has direct impact on accuracy. But as the layers increase, difficulty increases in training the huge network and also chance of introducing gradient vanishing issues increases. Hence, the drawbacks of traditional U-net based semantic segmentation can be removed by adding densely connected convolutional network with extra feature extractor to improve overall performance. The dense connection architecture is shown in Figure 1.

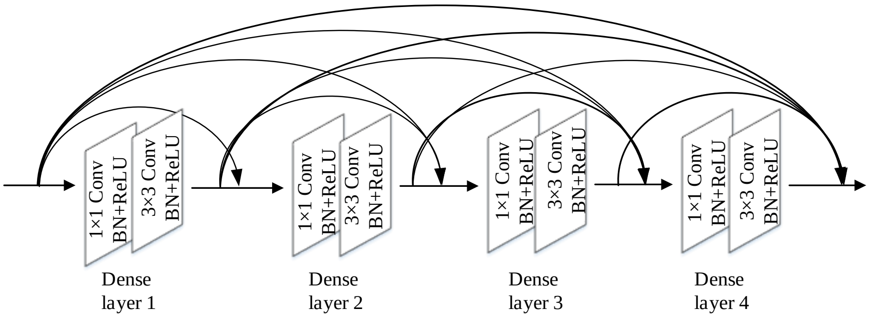


Figure 1. Dense connections with four dense layers

Dense connectivity reduces the overfitting problem for small dataset by proper regularization and reduced number of parameters by utilizing the feature reuse concept. In general, for dense connectivity, the l^{th} layer will receive all the featuremaps staged by the previous $(l-1)$ layers. Each layer is receiving a “collective knowledge” from all preceding layers which can be specified as

$$p_l = H_l([p_0, p_1, \dots, p_{l-1}])$$

where $(p_0, p_1, \dots, p_{l-1})$ describes combination of the output featuremaps in $(\mathrm{l}-1)$ layers. Here each layer l with function H_l produces k featuremaps where k is referred as growth rate parameter, which is generally a small value, so that the (l+1)th layer has $k \times (l-1) + k_0$ input featuremaps [34]. The growth rate k regulates the amount of information which can be added to each layer, network parameter space and performance.

Additionally, attention module is integrated with this dense U-net to make it learn the pixel-wise and channel-wise relationship over the images. The attention module consists of two modules namely the channel attention module (CAM) and the position attention module (PAM). The two attention blocks can reduce the redundancy among channels and focus on the most important parts of an image. It is implemented using sequential network (e.g. convolution layer, addition or multiplication) and activation function (e.g. a softmax or sigmoid) which adds a few trainable parameters to the network. These two modules shown in Figure 2 are explained below:

Let the input featuremap $\mathrm{K} \in \mathbb{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$ be feed into the convolution layer of attention module where C and W, H are number of channels and width and height of featuremap K. It generates two new featuremaps $\mathrm{L} \in \mathbb{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$ and $\mathrm{M} \in \mathbb{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$ that removes irrelevant features and edge information. Then, these two features are reshaped to $\mathbb{R}^{\mathrm{C}\times\mathrm{N}}$, where $\mathrm{N} = \mathrm{H} \times \mathrm{W}$ is number of pixels. The M transpose is multiplied with L and by applying a softmax function to produce $P \in R^{N\times N}$:

$$p_{ji} = \frac{exp\ (L_i \cdot M_j^T)}{\sum_{i=1}^{N} exp\ (L_i \cdot M_j^T)}$$

where, p_{ji} measures the i^{th}position effect on the j^{th}position. Later, input featuremap K is fed to convolution layer to generate a new featuremap $O \in \mathbb{R}^{C\times H\times W}$ which is reshaped as $\mathbb{R}^{C\times N}$. Then, the matrix multiplication between O and P is performed to produce a matrix of size $\mathbb{R}^{C\times N}$ and the result is reshaped as $\mathbb{R}^{C\times H\times W}$. At the end, the element-wise summation is performed with original feature K to obtain final output featuremap of size $\mathbb{R}^{C\times H\times W}$. This resulting featuremap is a weighted sum of features across all positions and original input features. Thus, it provides global contextual view and selectively aggregates long range contexts as per the spatial attention map [35]. Furthermore, the channel attention map is calculated based on input features $K \in \mathbb{R}^{C\times H\times W}$ by reshaping it to $\mathbb{R}^{C\times N}$ and then performing matrix multiplication between K and the transpose of K. Afterwards softmax function is applied to obtain channel attention map $X \in \mathbb{R}^{C\times C}$ which is expressed as:

$$x_{ji} = \frac{exp\ (K_i \cdot K_j^T)}{\sum_{i=1}^{C} exp\ (K_i \cdot K_j^T)}$$

where x_{ji} measures the i^{th}channel's impact on the j^{th} channel. Further, the matrix multiplication between X and K is performed and reshaped to $\mathbb{R}^{C\times H\times W}$. This channel map emphasizes on interdependent featuremapping and improves feature identification. Each channels' last feature is calculated by taking weighted sum of the all channels features and original features, which models the long-range semantic dependencies between featuremaps. CAM is able to highlight class-dependent featuremaps and discriminatively support a feature boost that cannot be produced by the convolution layers [35]. Finally, element-wise summation operation is performed on the outputs of channel and position attention module to obtain the final featuremap $T \in \mathbb{R}^{C\times H\times W}$.

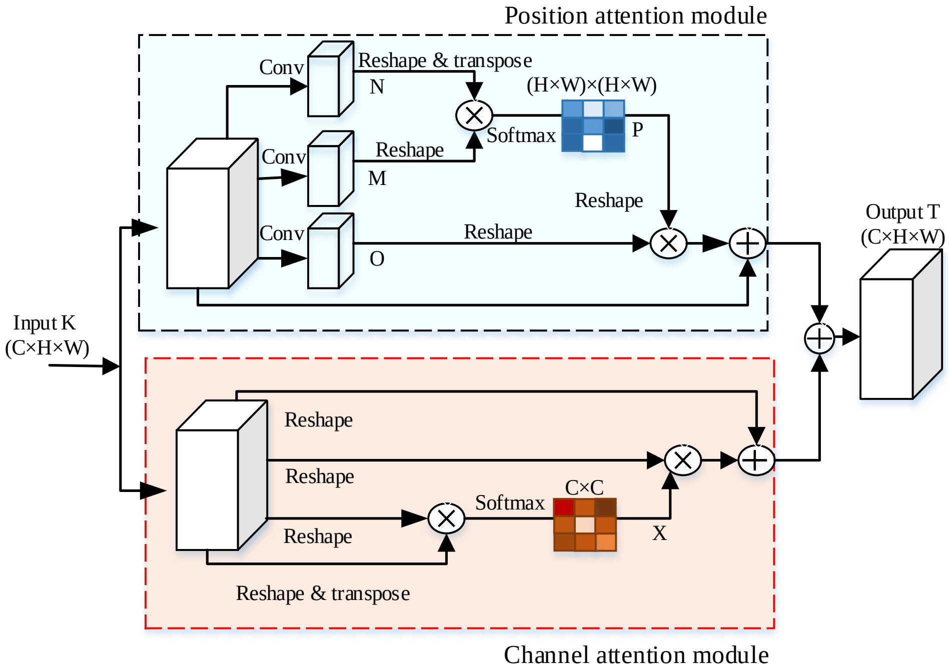


Figure 2. Detailed structure of position and channel attention module

In the proposed architecture, the main components are the encoder having dense block along with down-sampling, decoder consisting of dense block with up-sampling and skip connections between each layers of encoder decoder with attention module as in Figure 3. In the dense connection block, there is a set of two convolutional layers 1×1 and 3×3, likewise four sets are stacked together where each layer is concatenated with its preceeding layer as shown in Figure 1. This dense block has total 8 convolution layers which utilizes features extracted at each layer and maintain the resolution loss. Encoder part consists of four convolutional layers where each has different kernel size as 32, 64, 128 and 256 respectively. In each block number of features increases because of multiple dense connectivity. These can be reduced by using transition layer (downsampling). The downsampling is used for feature dimensionality reduction from previous dense layer connections. This layer has batch normalization and convolutional layer of 1×1 and lastly pooling layer. The number of channels reduces by two using 1×1 convolution while the size of the featuremap reduces by two using 2×2 average-pooling layers. In decoder part, the convolution layer is replaced by dense block while upsampling path is substantially like the U-Net mechanism. In the encoding path, each convolution is followed by a concatenation of features from corresponding layers. Lastly, the 1×1 convolution with sigmoid activation function is utilized to output the final pixel-wise classification. The overall performance of the U-Net technique is furthermore improved by utilising attention mechanism with dense connection module technique. Then the encoder side features are fed into two different attention blocks CAM and the PAM. The attention mechanism helps to improve the featuremap representation.

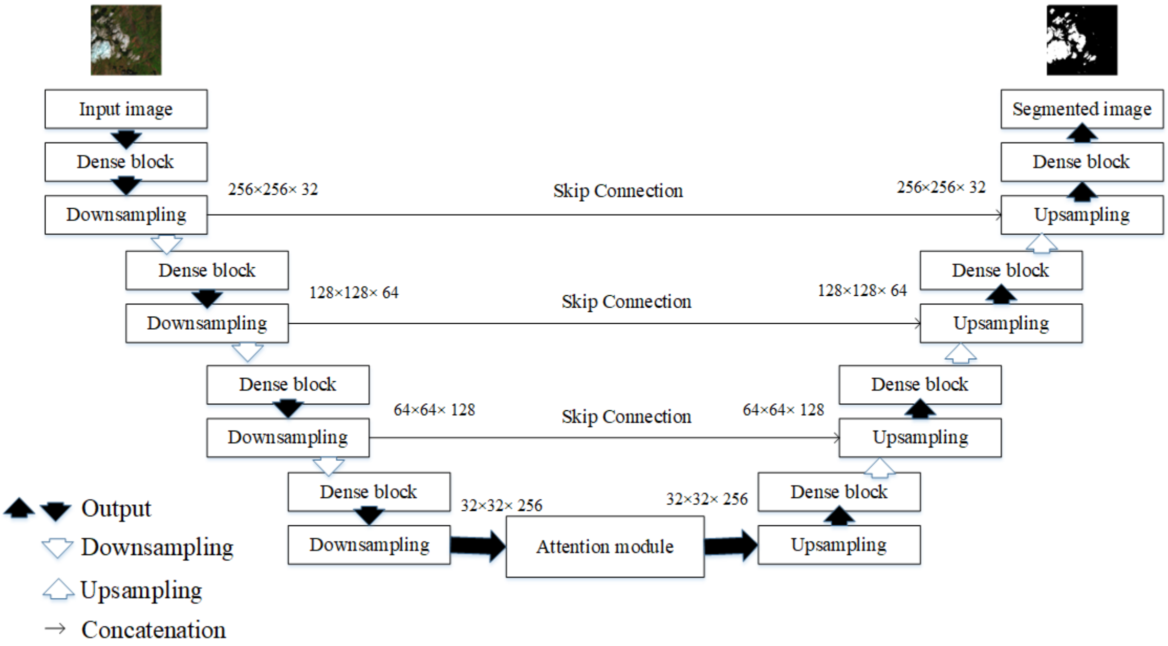


Figure 3. Implementation diagram of technique which consist of U-Net along with dense and attention module

3. RESULTS AND DISCUSSION

3.1. Dataset

The M. Joseph Hughes devolped SPARCS dataset in Oregon State University [28]. The Landsat 8 satellite acquired 80 images of 1000×1000 pixels with 11 spectral bands cloud data from OLI/TIRS sensor. It has different classes with manually generated mask for the different classes such as “snow/ice”,” water”, “cloud”, “cloud shadow”, and “flooded”. The dataset is downloaded from https://landsat.usgs.gov/spar:cs. Here the mask for broadly annotated into cloud and non-cloud class. The original dataset is splitted as 80% training and 20% for testing. Further the images are processed and cropped into 256×256 resulting into 1024 and 256 images are training and testing.

3.2. Evaluation metrics:

This section helps to compare the segmentation results which are being evaluated by the metrics which are precision, recall, F1-score and overall accuracy. The F1-score is the harmonic mean of recall and precision. These metrics helps to analyse and aid to improve the state-of-the-art technique.

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Overall\ accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

$$\text{F1} = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Where TP, TN, FP and FN are true positive, true negative, false positive and FN is false negative respectively for each class of label.

3.3. Training parameter details

The experiments were conducted on an NVIDIA GeForce GTX 1080 with 128 GB RAM. The main packages used are Python 3.6, CUDA 10.0, cuDNN 7.3 with keras and tensorflow library. The input image of size 256×256 is feed to the network. The hyper-parameters of the model play vital role and model accuracy is improve based on empirical hyperparameter values. The Adam optimizer is used with 16 batch size and 0.00001 learning rate. He initialiser [36] initializes all convolution kernel weights. Horizontal and vertical augmentation techniques helped to avoid overfitting. The sigmoid activation function is utilized to produce the final pixel values probability map. The loss function is binary cross entropy which is used to find the loss between ground truth image and predicted segmented output.

3.4. Numerical and visual results

Quantitative results obtained using attentional dense-U-Net architecture are tabulated in Table 1. These results show that the proposed technique can attain accurate segmentation results than some of state-of-the-art methods. From the experimental results using metrics such as precision, recall, f1-score and overall accuracy, it is clear that as epoch increases the accuracy increases. We tested for three different epochs 50, 100, and 200 separately which reveals that three architectures achieve a better result with 200 epochs as shown in Table 1. But with 100 epochs better result is achieved in a short time. First, we observe the results by utilizing simple U-Net architecture for semantic segmentation giving an accuracy of 93.18%, at the cost of overfitting. The dense connectivity helps to improve the performance by reducing the overfitting problem and improving the accuracy upto 94.87%. Dense connectivity requires significantly few parameters, encourage feature reuse and also does not have any performance degradation or overfitting. Lastly, three blocks namely U-net, dense module and attention module are integrated, so that the overall performance of the network is improved. The objective of proposed architecture is to extract more relevant features through multiple layers using dense connection as well as keep inter-spatial and inter-spectral information using attention block. The overall accuracy has improved to 96.02%. As in Table 1, three different architectures with its quantitative are tabulated. The U-Net inference time is 17.41, U-Net and Dense module takes 18.26 and combination of U-Net, Dense and Attention requires average time of 18.98 seconds.

Table 1. Results for detection of cloud using different proposed models based on U-Net techniques for different epochs

Models	Epoch	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
U-Net	50	84.13	81.36	81.89	86.78
U-Net+Dense module		88.47	87.01	87.13	89.12
U-Net+Dense+attention		89.76	87.45	89.24	91.40
U-Net	100	89.84	85.67	86.79	92.68
U-Net+Dense module		91.65	90.54	91.00	94.38
U-Net+Dense+attention		92.09	91.86	92.05	95.69
U-Net	200	90.46	86.74	87.47	93.18
U-Net+Dense module		92.87	91.48	92.42	94.87
U-Net+Dense+attention		93.14	92.71	93.00	96.02

The misclassified areas for different techniques are highlighted with red round circle as shown in Figure 4. It indicates that U-Net misidentifies small tiny cloud with non-cloud. U-Net with dense connectivity is able to remove artifacts, extract more accurate boundaries and hence a segmented image with higher quality is obtained. Further, incorporating attention module strengthens feature representation effectively. As shown in Figure 4(a) we have test images along with its groundtruth as in Figure 4(b). The results for U-Net, U-Net+dense module, and U-Net+dense+attention module is as shown in Figures 4(c)-(e) respectively. Figure 4(e) reveals better visualization results of image using attention module. With reduced misclassifications i.e clear boundaries between cloud and non-cloud pixels is the key benefits of the attention module. The cloud consisting of both thin and thick clouds are also identified properly. The result demonstrates that using attention module improves the performance of cloud and non-cloud discrimination.

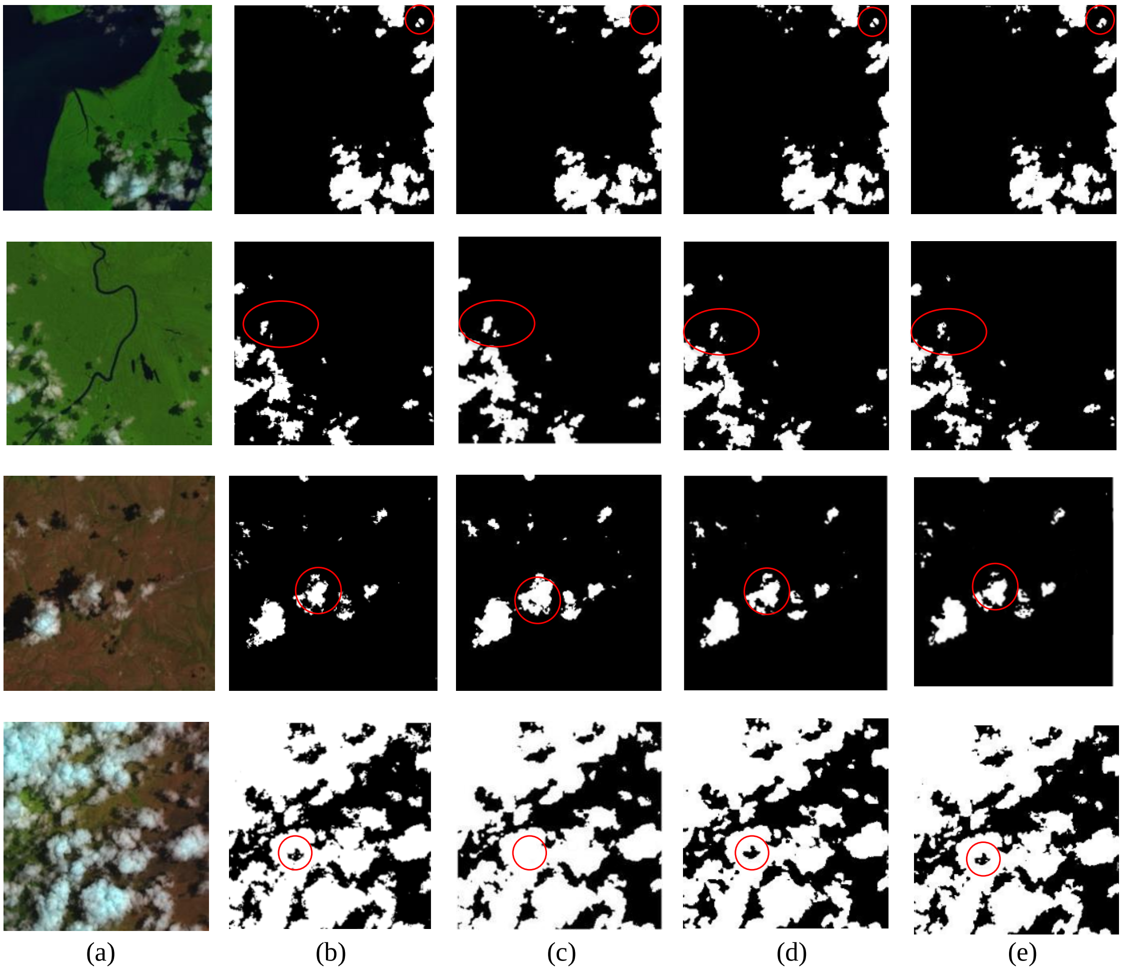


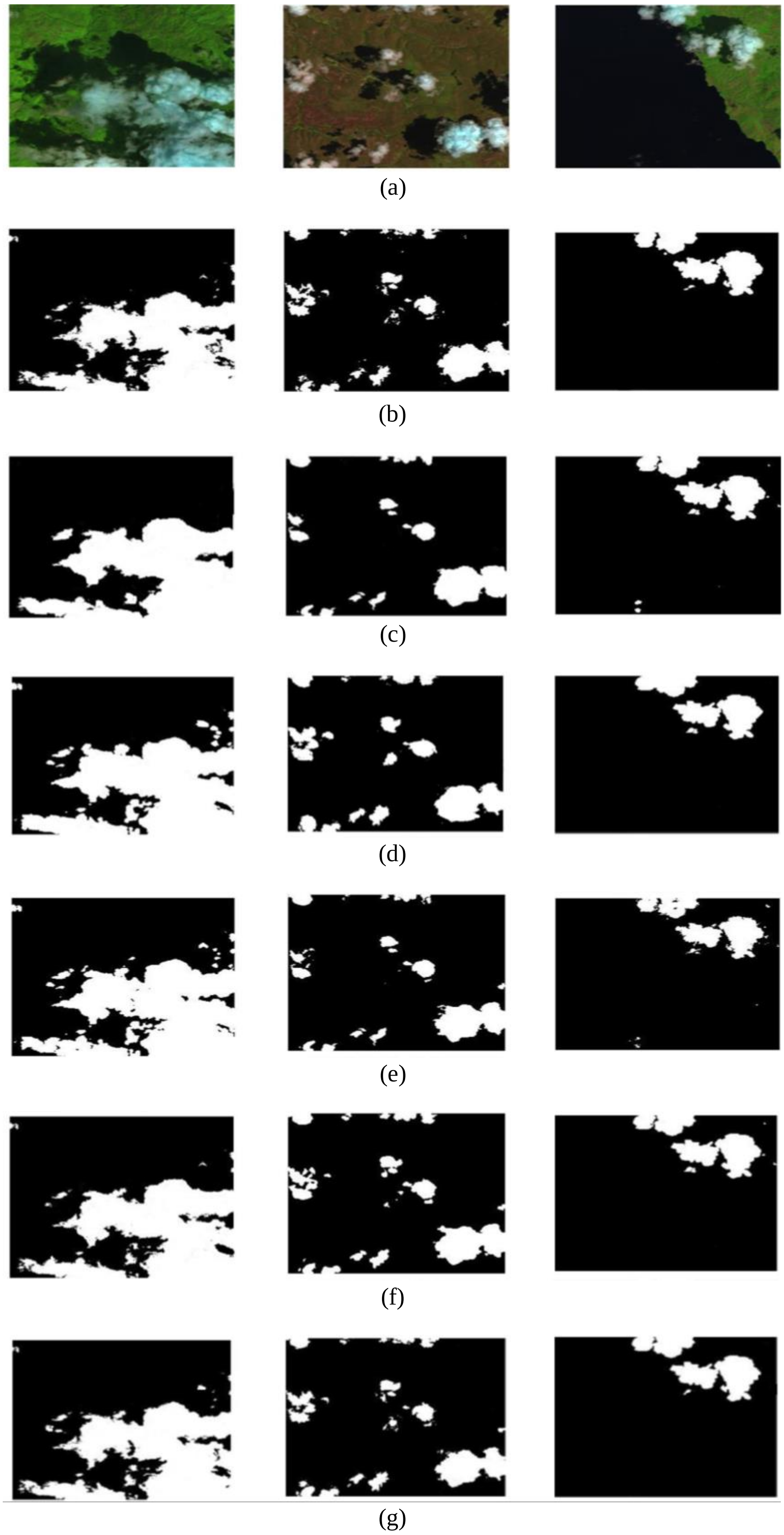
Figure 4. Remote sensing-based cloud detection visual analysis for different models (a) test images (b) groundtruth (c) U-Net (d) U-Net+dense module, and (e) U-Net+dense+attention module

3.5. Comparison with existing methods

We can observe the overall performance of all popular existing algorithms as compared with the proposed model of U-Net+dense net+attention module. These results are given in Table 2 and Figure 5. The various techniques compared are FCN, SegNet, multi-scale convolutional feature fusion (MSCFF) and CloudNet. Our implementations are performing better as compared to some of the existing methodologies such as FCN, SegNet and MSCFF. As shown in Figure 5(a) we have test images along with its groundtruth as in Figure 5(b). The results for FCN, SegNet, MSCFF and CloudNet are as shown in Figures 5(c)-(f). The modified U-Net model as in Figure 5(g) identifies the cloudy pixels more accurately with clear boundaries and has less chance of misclassifying tiny and thin cloudy pixels. The proposed technique is giving better results which are rich in multiscale information with clear and fine boundaries, while other techniques are failing in case of few less cloudy areas and finer boundaries.

Table 2. Comparison of different techniques along with metrics

Models	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
FCN [29]	85.48	87.45	85.21	87.17
SegNet [27]	81.54	89.01	90.40	90.29
MSCFF [37]	92.01	89.45	90.36	92.48
CloudNet [38]	90.46	88.48	90.14	94.01
U-Net+Dense+attention module	91.58	90.65	90.68	96.26



(a)

(b)

(c)

(d)

(e)

(f)

(g)

Figure 5. Various techniques results for cloud detection of remote sensing images (a) satellite image, (b) ground truth image, (c) FCN methodology, (d) SegNet methodology, (e) MSCFF methodology, (f) cloudNet methodology, and (g) proposed method

4. CONCLUSION

In these implementations, a U-Net architecture is utilized for detection of cloud to accurately segment cloudy and non-cloudy regions from high-resolution satellite images. Experiment performed on Landsat 8 images showed that the proposed techniques have good inference speed and less computational cost. We added dense connections to U-Net which boosted the reusabilty of the low-level features and maximization of the information flow between the layers. The attention module was incorporated into the existing architecture to enhance the productivity of semantic information dissemination by skip connections. It also adds only few parameters with stronger representative information, and the segmentation performance has improved effectively. The experimental accuracy of cloud detection significantly increased with dense connection and more improvement is observed with attention module. The proposed Dense-Net with attention module is able to achieve nearly 3% more accuracy compared to U-Net on SPARCS dataset. The overall analysis demonstrates that the proposed architecture can be utilized as a pre-processing step in remote sensing applications. As future work, modified attention module with residual connection can be utilized for better accurate segmentation.

REFERENCES

[1] H. Hashim, Z. A. Latif, and N. A. Adnan, "Land use land cover analysis with pixel-based classification approach," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 16, no. 3, pp. 1327–1333, Dec. 2019, doi: 10.11591/ijeecs.v16.i3.pp1327-1333.

[2] M. A. Zaytar and C. El Amrani, "Satellite image inpainting with deep generative adversarial neural networks," *IAES Int. J. Artif. Intell.*, vol. 10, no. 1, pp. 121–130, Mar. 2021, doi: 10.11591/ijai.v10.i1.pp121-130.

[3] J. D. J. H. Harikiran, "Hyperspectral image classification using support vector machines," *IAES Int. J. Artif. Intell.*, vol. 9, no. 4, p. 684, Dec. 2020, doi: 10.11591/ijai.v9.i4.pp684-690.

[4] K. M. Rao, B. S. Rao, B. S. Chandana, and J. Harikiran, "Dimensionality reduction and hierarchical clustering in framework for hyperspectral image segmentation," *Bull. Electr. Eng. Informatics*, vol. 8, no. 3, pp. 1081–1087, Sep. 2019, doi: 10.11591/eei.v8i3.1451.

[5] S. Y. J. Prasetyo, K. Dwi Hartomo, M. Chrismawati Paseleng, D. W. Chandra, and E. Winarko, "Satellite imagery and machine learning for aridity disaster classification using vegetation indices," *Bull. Electr. Eng. Informatics*, vol. 9, no. 3, pp. 1149–1158, Jun. 2020, doi: 10.11591/eei.v9i3.1916.

[6] N. Ismail and K. Nizam Tahar, "Semi-automatic building footprint using multirotor and fixed wing UAV," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 17, no. 3, pp. 1298–1305, Mar. 2020, doi: 10.11591/ijeecs.v17.i3.pp1298-1305.

[7] M. Dimyati, K. Kustiyo, and R. D. Dimyati, "Paddy field classification with MODIS-terra multi-temporal image transformation using phenological approach in Java Island," *Int. J. Electr. Comput. Eng.*, vol. 9, no. 2, pp. 1346–1358, Apr. 2019, doi: 10.11591/ijece.v9i2.pp1346-1358.

[8] D. P. Y. Suseno and T. J. Yamada, "Two-dimensional, threshold-based cloud type classification using MTSAT data," *Remote Sens. Lett.*, vol. 3, no. 8, pp. 737–746, Dec. 2012, doi: 10.1080/2150704X.2012.698320.

[9] A. Mefti, A. Adane, and M. Y. Bouroubi, "Satellite approach based on cloud cover classification: Estimation of hourly global solar radiation from meteosat images," *Energy Convers. Manag.*, vol. 49, no. 4, pp. 652–659, Apr. 2008, doi: 10.1016/j.enconman.2007.07.041.

[10] A. A. K. Tahir, "A system based on ratio images and quick probabilistic neural network for continuous cloud classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 12, pp. 5008–5015, Dec. 2011, doi: 10.1109/TGRS.2011.2153863.

[11] Q. Li, W. Lu, and J. Yang, "A hybrid thresholding algorithm for cloud detection on ground-based color images," *J. Atmos. Ocean. Technol.*, vol. 28, no. 10, pp. 1286–1296, Oct. 2011, doi: 10.1175/JTECH-D-11-00009.1.

[12] O. Kärner, "A multi-dimensional histogram technique for cloud classification," *Int. J. Remote Sens.*, vol. 21, no. 12, pp. 2463–2478, Jan. 2000, doi: 10.1080/01431160050030565.

[13] T. A. Berendes, J. R. Mecikalski, W. M. MacKenzie, K. M. Bedka, and U. S. Nair, "Convective cloud identification and classification in daytime satellite imagery using standard deviation limited adaptive clustering," *J. Geophys. Res.*, vol. 113, no. D20, p. D20207, Oct. 2008, doi: 10.1029/2008JD010287.

[14] Z. Ameur, S. Ameur, A. Adane, H. Sauvageot, and K. Bara, "Cloud classification using the textural features of Meteosat images," *Int. J. Remote Sens.*, vol. 25, no. 21, pp. 4491–4503, Nov. 2004, doi: 10.1080/01431160410001735120.

[15] T. Inoue, "A cloud type classification with NOAA 7 split-window measurements," *J. Geophys. Res.*, vol. 92, no. D4, p. 3991, 1987, doi: 10.1029/JD092iD04p03991.

[16] U. Amato *et al.*, "Statistical cloud detection from SEVIRI multispectral images," *Remote Sens. Environ.*, vol. 112, no. 3, pp. 750–766, Mar. 2008, doi: 10.1016/j.rse.2007.06.004.

[17] H.-Y. Cheng and C.-C. Yu, "Block-based cloud classification with statistical features and distribution of local texture features," *Atmos. Meas. Tech.*, vol. 8, no. 3, pp. 1173–1182, Mar. 2015, doi: 10.5194/amt-8-1173-2015.

[18] S. A. Ackerman, K. I. Strabala, W. P. Menzel, R. A. Frey, C. C. Moeller, and L. E. Gumley, "Discriminating clear sky from clouds with MODIS," *J. Geophys. Res. Atmos.*, vol. 103, no. D24, pp. 32141–32157, Dec. 1998, doi: 10.1029/1998JD200032.

[19] Z. Zhu, S. Wang, and C. E. Woodcock, "Improvement and expansion of the Fmask algorithm: cloud, cloud shadow, and snow detection for Landsats 4-7, 8, and Sentinel 2 images," *Remote Sens. Environ.*, vol. 159, pp. 269–277, Mar. 2015, doi: 10.1016/j.rse.2014.12.014.

[20] Y. Zhang, B. Guindon, and J. Cihlar, "An image transform to characterize and compensate for spatial variations in thin cloud contamination of Landsat images," *Remote Sens. Environ.*, vol. 82, no. 2–3, pp. 173–187, Oct. 2002, doi: 10.1016/S0034-4257(02)00034-2.

[21] Y. Yuan and X. Hu, "Bag-of-words and object-based classification for cloud extraction from satellite imagery," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 8, no. 8, pp. 4197–4205, Aug. 2015, doi: 10.1109/JSTARS.2015.2431676.

[22] M. Shi, F. Xie, Y. Zi, and J. Yin, "Cloud detection of remote sensing images by deep learning," in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Jul. 2016, vol. 2016-Novem, pp. 701–704, doi: 10.1109/IGARSS.2016.7729176.

[23] F. Xie, M. Shi, Z. Shi, J. Yin, and D. Zhao, "Multilevel cloud detection in remote sensing images based on deep learning," *IEEE*

J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 10, no. 8, pp. 3631–3640, Aug. 2017, doi: 10.1109/JSTARS.2017.2686488.

[24] Y. Chen, R. Fan, M. Bilal, X. Yang, J. Wang, and W. Li, "Multilevel cloud detection for high-resolution remote sensing imagery using multiple convolutional neural networks," *ISPRS Int. J. Geo-Information*, vol. 7, no. 5, p. 181, May 2018, doi: 10.3390/ijgi7050181.

[25] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," vol. 7, pp. 43369–43382, Nov. 2014, [Online]. Available: http://arxiv.org/abs/1411.4038.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9351, 2015, pp. 234–241.

[27] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: a deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.

[28] Y. Zhan, J. Wang, J. Shi, G. Cheng, L. Yao, and W. Sun, "Distinguishing cloud and snow in satellite images via deep convolutional network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1785–1789, Oct. 2017, doi: 10.1109/LGRS.2017.2735801.

[29] S. Mohajerani, T. A. Krammer, and P. Saeedi, "A cloud detection algorithm for remote sensing images using fully convolutional neural networks," in *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, Aug. 2018, pp. 1–5, doi: 10.1109/MMSP.2018.8547095.

[30] J. Lu *et al.*, "P_Segnet and NP_Segnet: New Neural Network Architectures for Cloud Recognition of Remote Sensing Images," *IEEE Access*, vol. 7, pp. 87323–87333, 2019, doi: 10.1109/ACCESS.2019.2925565.

[31] Z. Zhang, A. Iwasaki, G. Xu, and J. Song, "Small satellite cloud detection based on deep learning and image compression," *Preprints*, no. February, pp. 1–12, 2018.

[32] J. Yang, J. Guo, H. Yue, Z. Liu, H. Hu, and K. Li, "CDnet: CNN-based cloud detection for remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 6195–6211, Aug. 2019, doi: 10.1109/TGRS.2019.2904868.

[33] J. H. Jeppesen, R. H. Jacobsen, F. Inceoglu, and T. S. Toftegaard, "A cloud detection algorithm for satellite imagery based on deep learning," *Remote Sens. Environ.*, vol. 229, no. May, pp. 247–259, Aug. 2019, doi: 10.1016/j.rse.2019.03.039.

[34] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, vol. 2017-Janua, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.

[35] J. Fu *et al.*, "Dual attention network for scene segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, vol. 2019-June, pp. 3141–3149, doi: 10.1109/CVPR.2019.00326.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: surpassing human-level performance on imagenet classification," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, vol. 2015 Inter, pp. 1026–1034, doi: 10.1109/ICCV.2015.123.

[37] Z. Li, H. Shen, Q. Cheng, Y. Liu, S. You, and Z. He, "Deep learning based cloud detection for medium and high resolution remote sensing images of different sensors," *ISPRS J. Photogramm. Remote Sens.*, vol. 150, pp. 197–212, Apr. 2019, doi: 10.1016/j.isprsjprs.2019.02.017.

[38] C.-C. Liu *et al.*, "Clouds classification from sentinel-2 imagery with deep residual learning and semantic image segmentation," *Remote Sens.*, vol. 11, no. 2, p. 119, Jan. 2019, doi: 10.3390/rs11020119.

BIOGRAPHIES OF AUTHORS

Aarti Kumthekar received B. E in Electronics and Communication Engineering in 2011 and completed M.E in 2013 from BVCOE, Kolhapur. She is currently doing her research in VIT Vellore. Her research interest Image processing, Satellite Image processing, Machine learning and Deep learning. She can be contacted at email: aartikumthekar@gmail.com.

Gudheti Ramachandra Reddy received the M.Sc. and M.Sc. (Tech.) degrees from the Birla Institute of Technology and Science, Pilani, India, in 1973 and 1975, respectively, and the Ph.D. degree from the Indian Institute of Technology Madras, Chennai, India, in 1987. From 1976 to 2010, he was with the Department of Electrical and Electronics Engineering, College of Engineering, Sri Venkateswara University, Tirupati India. From February 1989 to May 1991, he was Visiting Scientist with the Department of Electrical and Computer Engineering, Concordia University, Montreal, QC, Canada. Currently, he is a senior Professor of School of Electronics Engineering, Vellore Institute of Technology, Vellore, India. Prof. Reddy is a Fellow of the Institution of Electronics and Telecommunication Engineers and a member of the Indian Society for Technical Education. He can be contacted at email: grreddy@vit.ac.in.