Journal on
Artificial Intelligence

Tech Science Press

DOI: 10.32604/jai.2022.034931
Article

ResCD-FCN: Semantic Scene Change Detection Using Deep Neural Networks

S. Eliza Femi Sherley[1,*], J. M. Karthikeyan[1], N. Bharath Raj[1], R. Prabakaran[2], A. Abinaya[1] and S. V. V. Lakshmi[3]

[1]Department of Information Technology, Anna University, MIT Campus, Chennai, 600044, India
[2]Computer Center, Anna University, MIT Campus, Chennai, 600044, India
[3]Department of Computer Science and Engineering, Anna University, CEG Campus, Chennai, 600025, India
*Corresponding Author: S. Eliza Femi Sherley. Email: selizafemisherley@gmail.com
Received: 01 August 2022; Accepted: 23 September 2022; Published: 24 May 2023

Abstract: Semantic change detection is extension of change detection task in which it is not only used to identify the changed regions but also to analyze the land area semantic (labels/categories) details before and after the timelines are analyzed. Periodical land change analysis is used for many real time applications for valuation purposes. Majority of the research works are focused on Convolutional Neural Networks (CNN) which tries to analyze changes alone. Semantic information of changes appears to be missing, there by absence of communication between the different semantic timelines and changes detected over the region happens. To overcome this limitation, a CNN network is proposed incorporating the Resnet-34 pre-trained model on Fully Convolutional Network (FCN) blocks for exploring the temporal data of satellite images in different timelines and change map between these two timelines are analyzed. Further this model achieves better results by analyzing the semantic information between the timelines and based on localized information collected from skip connections which help in generating a better change map with the categories that might have changed over a land area across timelines. Proposed model effectively examines the semantic changes such as from-to changes on land over time period. The experimental results on SECOND (Semantic Change detectiON Dataset) indicates that the proposed model yields notable improvement in performance when it is compared with the existing approaches and this also improves the semantic segmentation task on images over different timelines and the changed areas of land area across timelines.

Keywords: Remote sensing; convolutional neural network; semantic segmentation; change detection; semantic change detection; resnet; FCN

1 Introduction

Change detection (CD) task involves detecting the changes observed in land areas using satellite images between given time intervals. CD task is very useful for many real-world applications like

planning and monitoring urban growth, managing resources pertaining to the urban areas, monitoring environment activities, damage assessment after natural disasters and in many such social applications. Even though the binary CD models allows the users to monitor and evaluate the regions of interest in satellite images, the information observed is small and it doesn't describe in elaborated manner. Most of the CD applications tend to know what changes happened other than where the changes could have been happened. To overcome this scenario, semantic change detection (SCD) task is defined [1–7] which provides change information [8,9] and detailed land change maps which includes change information corresponding to each category. Recently with the development of CNN [10], massive attainment has been achieved in the field of CD task. The CNNs learn to segment each category based on the scene and generate SCD maps. CNNs have a bottom-up architecture design in which the features of bi-temporal region [11, 12] are merged and downscaled through the stack of convolutional layer blocks. The details of changes are recognized through the weights that have been computed and feature transformations. In comparison with traditional approaches, CNN-based approaches tend to be advantageous with features like improved robustness, effective to model more complex changes, etc. With the interpretation of image processing, Binary Change Detection (BCD) task is known a binary segmentation method where binary change map is produced to indicate the changed/unchanged regions of interest. The conventional CNN-based methods don't fit for SCD task since they are single-ended in nature producing the CD map alone [13]. To fill the research gap in order to carry out SCD task, a novel CNN-based model is been proposed. As a part of proposed methodology to perform SCD task, two sub tasks of SCD are carried out: Semantic Segmentation (SS) of the land regions and CD of the land regions in which the extracted features are shared and fused together. The loss functions are used separately to monitor the SS and CD tasks of the SCD. The rest of this paper is organized as follows. Section 2 presents the related works of CD carried out by processing the satellite images. Section 3 & 4 explains the system architecture and implementation of proposed novel CNN model. Section 5 describes the various experimental settings used in this study and the evaluation metrics used and lists the results of the experiments. Section VI draws the conclusion.

2 Related Works

This section is presented with various approaches and the developments in CD tasks which are in practice over years. It also comprises learning about how CNN-based methods are introduced in CD approaches and recent methods used for SCD are reviewed.

Tomoyuki Suzuki et al. (2018) [14] have proposed a method to identify the semantic meaning from changed regions. Their work mainly focused on semantic segmentation, along with the traditional approaches to perform change detection. To achieve good performance, improvements are done to hyper column representations which is commonly referred as hypermaps, using convolutional maps derived from convolutional neural networks (CNNs). Image patches are processed to extract multi-scale features which helps for the process of SCD. Tsunami Panorama Change Detection (TSUNAMI) dataset is processed in which the modified parts are re-annotated through semantic classes. Problems associated with those processes are usually because of the result of light source fluctuations and changes in viewing angle.

Yuan et al. (2020) [15] in their work presented a pre-training technique under self-supervision for initializing transformer based networks. Satellite Image Time Series (SITS) data is processed to represent spectral-temporal information of land cover semantics. Pre-trained techniques which are self–supervised in nature helps to address the problem of poorly labeled samples. Alternate approach

to convolution and recurrent neural networks is introduced for classifying SITS data. This reduces the risk of overfitting and also improves the model performance.

Yang et al. (2022) [16] in their work presented an Asymmetric Siamese Network (ASN) which extracts pair of features from areas which is of different size to locate and identify semantic changes. Model training and evaluation is improvised by introducing an adaptive threshold learning technique which is tested using SECOND dataset. Impact of the model is evaluated using Segregated Kappa (SeK) coefficient, where results show that the model consistently achieve better results.

Peng et al. (2021) [17] have proposed Siamese U-Net architecture based convolutional network to perform large-scale SCD (SCDNet). With the encoder-decoder structure, additional unit is added at the end of the encoder which exploits multi-scale information. Semantic change map which gets generated as an output includes both binary and semantic change detection information. Attention mechanism and monitoring strategies are incorporated in this network model to improve its performance.

Sun et al. (2022) [18] have presented a technique with Conv-LSTM which provides end-to-end spatiotemporal network, thereby it process both spatial and temporal information. Convolution and recurrent structures are combined in a single layer which helps to improvise the model performance. Experimentation is carried out with SZTAKI and Beichuan datasets.

Wang et al. (2022) [19] in their research proposed a densely connected functional aggregation module (DCFAM) with the SwinTransformer as its base to extract multi-scale relationship information. Decoder part of DCFAM module extracts contextual information and helps to restore resolution of the images and generate accurate segmentation maps. Experiments are carried out on ISPRS Vaihingen and Potsdam dataset to evaluate this model, thereby it is concluded that this model performs better for segmentation task.

Hua et al. (2022) [20] have presented a framework to segment semantic information from aerial images with incomplete annotations, where few pixels are annotated using easy-to-draw scribbles. With the limited annotations drawn, FEature and Spatial relational regulArization (FESTA) method helps to perform supervised task along with unsupervised learning signals by which spatial information and features of neighborhood structures are analyzed. Numerically and visually results shows that this method provide better results in segmenting semantic information.

In order to summarize the review of existing works, it can be concluded that the existing approaches gave an outline to understand about the various approaches used to process the semantic scene changes between two timeframe. Existing models faced the following drawbacks when performing SCD tasks: Few existing models were unable to retain significant convolutional features and extract complex features, resulting in reduced performance and inaccurate results. Existing approaches' high model complexity and lack of attention mechanisms limit the performance of SCD tasks in real time. Fluctuations in the light source, as well as a lack of labels in the dataset, limit the performance in analysing semantic changes over time. The studies have helped to understand the underlying disadvantages and advantages of the existing methods and it also helped to build a novel architecture which overcome the limitations of the existing system by incorporating a ResNet based architecture with FCN blocks to extract more features and this also supports with the localized changes which relates between two timeframes. Hence, it produces the semantic scene change maps with precise results.

3 Proposed ResCD-FCN Model for SCD

The SCD task and its dependent subtasks such as Semantic Segmentation (SS) and Binary Change Detection (BCD) tasks can be defined based on a given input image M, where SS does the task to find function (q_s) that maps the image M into a semantic map O (see Eq. (1)).

$$q_s\left(m_{i,j}\right) = k_{i,j} \tag{1}$$

where $m_{i,j}$ denotes a pixel on M, $k_{i,j}$ is estimated as semantic class of satellite images. The BCD task q_{bcd} evaluates two images M_1, M_2 [9,12] into a change map O (see Eq. (2)). The image pixels such as $m1_{i,j}$, $m2_{i,j}$ on M_1, M_2 [9,12] are related to the same region, the calculation for this is given as follows:

$$q_{bcd}\left(m1, m2\right) = \begin{cases} 0, & k1 = k2 \\ 1, & k1 \neq k2 \end{cases} \quad for\ each\ i,j \tag{2}$$

where the signal estimated lists whether there is a change in the satellite image semantic classes or not. The SCD function q_{scd} is a union of q_s and q_{bcd} (see Eq. (3)):

$$q_{scd}\left(m1, m2\right) = \begin{cases} (0,0), & k1 = k2 \\ (k1, k2), & k1 \neq k2 \end{cases} \quad for\ each\ i,j \tag{3}$$

The result q_{scd} provides two semantic change maps O_1 and O_2 which indicates the change location and semantic classes of satellite images for the given region. The ground truth of the SCD tasks includes G_1 and G_2. In this proposed ResCD-FCN model, two FCN-CNN encoders T_1 and T_2 are used to extract the semantic information from M_1 and M_2. The extracted semantic features are combined to train ResCD module (D), which draws the information about the difference between regions. The calculation can be represented as follows (see Eqs. (4)–(6)):

$$L_1, L_2 = T_1\left(M_1\right), T_2\left(M_2\right) \tag{4}$$

$$O = D\left[T_1\left(M_1\right), T_2\left(M_2\right)\right] \tag{5}$$

$$K_1, K_2 = O.\left(L_1, L_2\right) \tag{6}$$

The function of FCN block is to extract features from the satellite images based on down-sampling technique with more number of blocks which includes Convolution layer and BatchNorm layer. The FCN block is constructed with dense layers in which the ResNet-34 pretrained model is added to fine-tune its performance. The layers of pre-trained ResNet-34 model are used in constructing the FCN block which consists of five layers as specified in the figure (see Fig. 4).

The head layer consists of a collection of convolutional layers with BatchNorm and ReLU activation layers. Here, weights and bias of the model are initialized based on the convolutional layer/linear layer or BatchNorm layer which is used in architecture of the model.

The ResCD block consists of a series of convolutional and BatchNorm layers in which we tend to extract features down the way from the model and the model gets appended with the initial input feature map as a skip connection which helps in identifying the localized features and ReLU activation function is applied to get the corresponding result. The main use of ResCD-FCN block is to extract maximum number of features with localized details from ResCD block and to extract features in more detailed manner from FCN block. So, this process helps to identify the optimized features which basically assist in building a better scene change detection map.

The two satellite images are fed into ResCD-FCN block in which each satellite image will be processed by FCN block independently to extract the features. The change detection map can be identified based on concatenation of the extracted features of two satellite images and then fed into ResCD-FCN block to get the localized information of the changes that may have happened between the years and finally a collection of Convolutional and BatchNorm layers are used in constructing the change map.

The semantic analysis of each year can be analyzed using a convolutional layer with the number of classes as its output labels. Based on the maps created, up sampling is applied on the images using bilinear interpolation technique to the required output dimension. The advantages of having ResCD-FCN block includes 1) Helps in interpreting Semantic class and change information explicitly. 2) The model perceives semantic changes in detailed manner using the features extracted from the temporal partitions.

3.1 Loss Functions

Two loss functions are used while training the ResCD-FCN model, the semantic class related loss W_s, the binary change related loss W_c. The semantic loss W_s represents the multiclass cross entropy loss between the semantic segmentation results O_1, O_2 and the ground truth semantic change maps G_1,G_2. The calculation of W_s is as follows (see Eq. (7)):

$$W_s = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(m_i) \tag{7}$$

where N represents the number of classes as mentioned in the dataset which is used, y_i and m_i denote the original Ground Truth and predictions of i^{th} class.

The change loss W_c (see Eq. (8)) represents the binary-cross entropy loss between the change map G_c and predicted change map O. The G_c is produced with G_1 or G_2 by replacing the non-changed labels with changed labels. The W_c is calculated as:

$$W_c = -y_c \log(m_c) - (1 - y_c)\log(1 - m_c) \tag{8}$$

where y_c denotes the ground truth and m_c denotes the prediction probability with respective to the change. The training of feature partition blocks is directly dependent by G_1 and G_2 and is supported by G_c while training is carried out to perform CD task which is dependent by G_c.

The relationship between the three outputs M_1,M_2,O and the ground truth maps G_1,G_2 and G_c are given by the total loss W_{scd}, which is indicated as (see Eq. (9)):

$$W_{scd} = (W_{m1} + W_{m2})/2 + W_c \tag{9}$$

where, W_{m1}, W_{m2} represents the semantic loss information which corresponds to each temporal partition respectively. Finally, they are summed up and averaged to represent W_s.

3.2 System Architecture

The following figures (See Figs. 1–4) represents the architecture of the proposed model. The ResCD-FCN model (See Fig. 1) is made up of FCN blocks (See Fig. 4) that take in satellite images and extract features before feeding them into the ResCD block, where the extracted features are used to

calculate the difference between the regions. The main purpose of the ResCD-FCN block (See Fig. 2) is to extract as many features as possible with localized details from the ResCD block (See Fig. 3) and to extract features in a more detailed manner from the FCN block. As a result, this enables us to recognize the optimized features that aid in the development of an improved scene change detection map. The model's classifier block generates the semantic map of T1 and T2 images, while the CD block generates the overall semantic map of T1 and T2 with differences between regions.

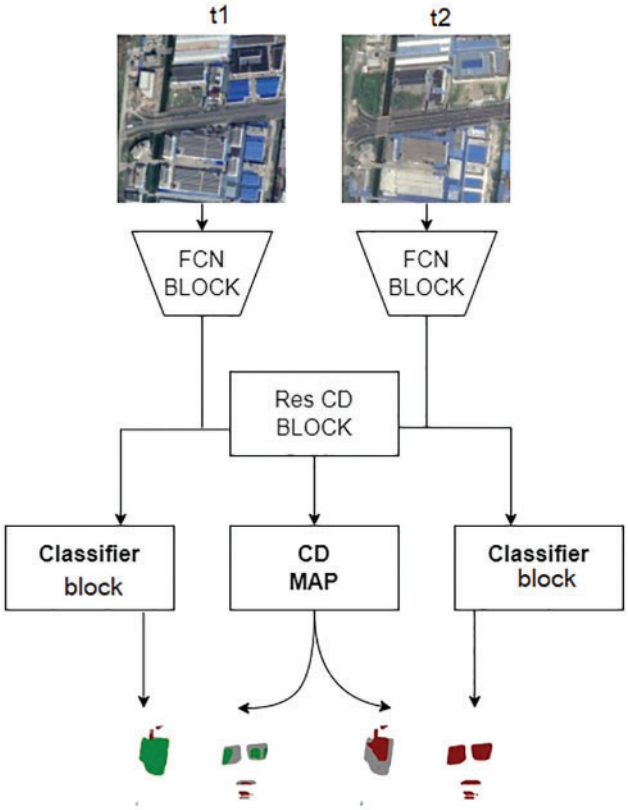

Figure 1: Overview of ResCD-FCN

4 Implementation

The Semantic Change Detection Dataset (SECOND), is mapped to the number of classes as mentioned in the dataset and each class is assigned with a unique colour corresponding to each label. The pre-processing of training data includes data augmentation techniques such as Random Flip and Random Rotation applied on the satellite images and also for the labels of two years in which all the images are flipped based on a certain threshold. Then the augmented images are read in a shuffled manner. The testing data of dataset includes satellite images of two years and it is read in non-shuffled way.

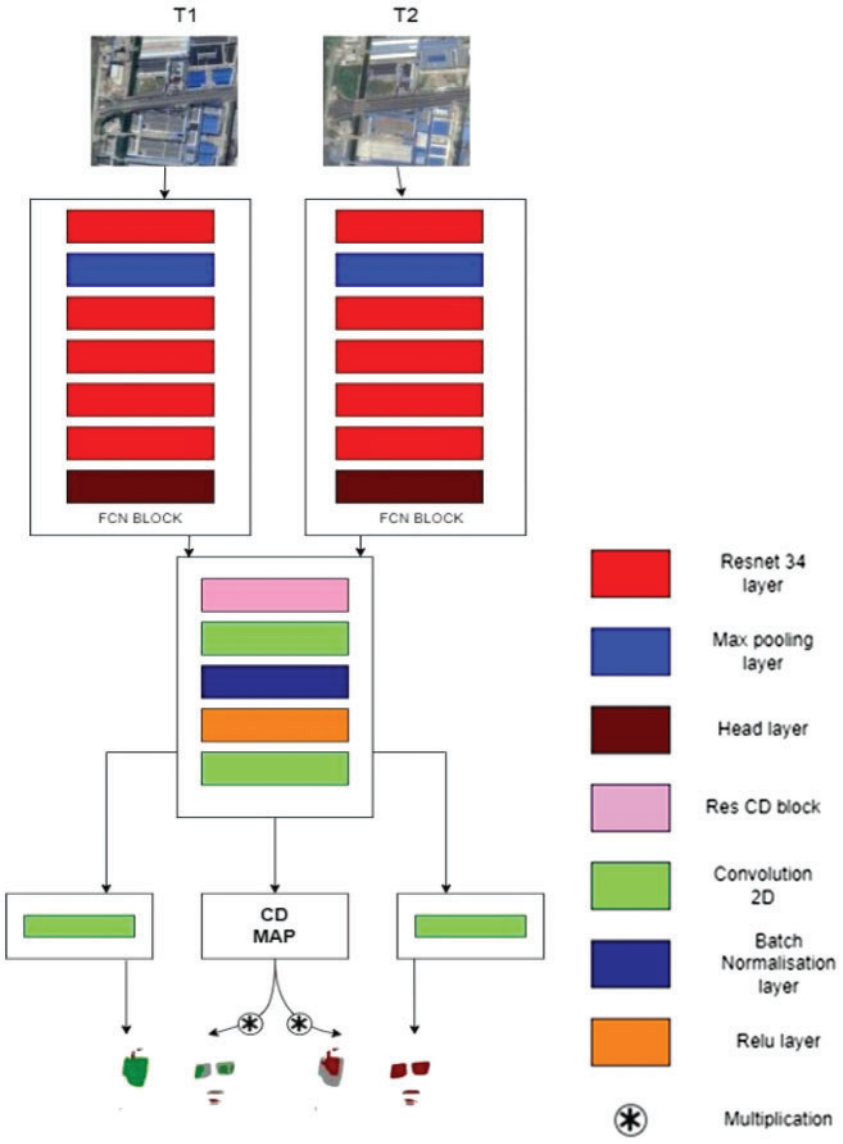

Figure 2: Detailed architecture of ResCD-FCN

The model is trained by feeding the satellite images and its labels of two years and backward propagation is done to ensure that the model is trained by the optimum weights by adjusting its learning rate, momentum and weight decay based on addition of half the value of Cross Entropy loss function the outputs of first year and its corresponding ground truth label and outputs of second year and its corresponding ground truth label. The Stochastic gradient descent (SGD) optimizer is applied to adjust the training process in order to extract the optimum weights. The Binary cross entropy (BCE) loss function is applied on the summative loss obtained then this is compared with the change detection map generated from the model. The Change Similarity criterion is applied on the outputs of first year, second year and output from labels in which the difference between the targets, that is the changed image and unchanged image is computed in order to obtain more accurate prediction. Finally, all the losses are summed up together and then back propagation is applied which helps in generalizing the

model. The change map is generated as an output of the proposed model which identifies the changes using thresholding approach. With threshold greater than 0.5, the pixel is assigned as it has a changed region else it remains unchanged. This in turn is compared with the labels to reduce the loss developed in training process.

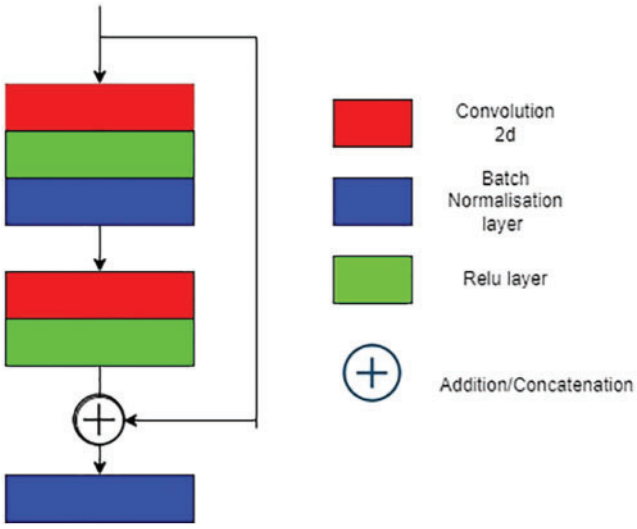

Figure 3: ResCD component

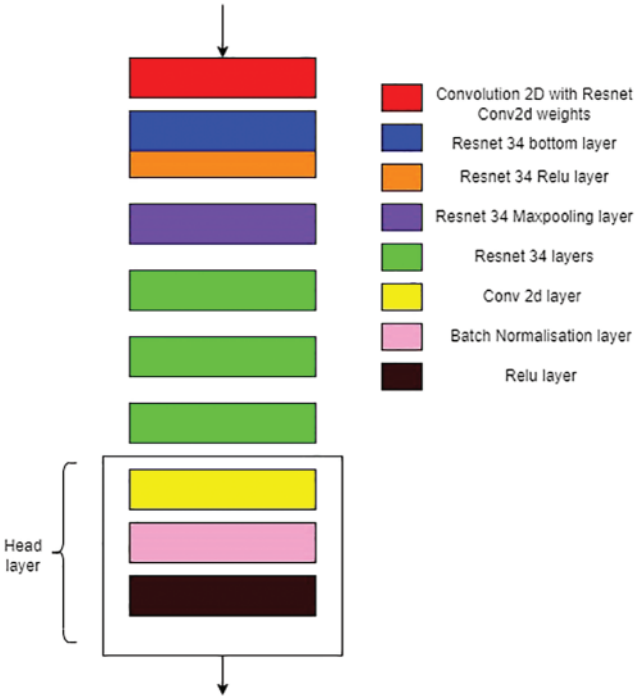

Figure 4: FCN component

The model is tested by feeding the satellite image features of two years. The loss is calculated based on addition of half of Cross Entropy loss function, with the outputs of first year and its ground truth label and outputs of second year and its ground truth label. The change map is generated and the results identified from the model are either pixel is assigned to a changed region or it's assigned as unchanged region, also it is compared with the ground truth label to quantify and assess the metrics of the validation process. The predictions of the bi-temporal semantic labels are generated by multiplying each timeline individually with the change map and it gets stored.

5 Dataset Description and Experimental Settings

5.1 Dataset

The experiments in this work is carried out in the Semantic Change Detection Dataset (SECOND) [12], a benchmark dataset to perform SCD. The SECOND dataset is well-annotated and it includes HR optical (RGB channels) images of two years collected from a collection of aerial platforms and sensors. The pair of images covers the regions in China, including Shanghai, Chengdu, Hangzhou, etc. Each image has 512x512 dimensions and is annotated at the pixel level. The semantic class type of the changed locations is also provided along with the dataset. It includes six land cover classes which contributes to analyze the natural and man-made changes. This dataset includes the 2968 pairs of images, its split into training and testing set with 2375 image pairs for training, 593 pairs for testing.

5.2 Evaluation Metrics

Three metrics are used to evaluate the proposed SCD model. Tasks in SCD are tested using metrics like overall accuracy (OA), mean Intersection over Union (MIoU) and Separated Kappa (SeK) coefficient. The confusion matrix is generated as, $A = \{a_{m,n}\}$ where $a_{m,n}$ denotes the number of pixels that are classified as 'm' and the ground truth index is denoted as 'n' (m, n ∈ {0, 1, …., N}). OA is computed as (see Eq. (10))

$$OA = \sum_{m=0}^{N} a_{mm} \Big/ \sum_{m=0}^{N} \sum_{n=0}^{N} a_{mn} \tag{10}$$

Since OA is calculated by identifying the 'no change' pixels, this metric alone is insufficient to evaluate the semantic segmentation of changed classes. The MIoU and SeK are the other two metrics which assist to evaluate the 'change/no-change' regions and also semantic segmentation of classes which are under 'change' category respectively. MIoU (see Eq. (11)) denotes the mean value of IoU that is IoU of no-change regions (IoU_{nc}) and IoU of the changed regions (IoU_c) (see Eq. (12), Eq. (13)):

$$MIoU = (IoU_{nc} + IoU_c) \Big/ 2 \tag{11}$$

$$IoU_{nc} = a_{00} \Big/ \left(\sum_{m=0}^{N} a_{m0} + \sum_{n-0}^{N} a_{0n} - a_{00} \right) \tag{12}$$

$$IoU_c = \sum_{m=1}^{N} \sum_{n=1}^{N} a_{mn} \Big/ \left(\sum_{m=0}^{N} \sum_{n=0}^{N} a_{mn} - a_{00} \right) \tag{13}$$

The SeK coefficient is calculated based on the confusion matrix $A' = \{a_{mn}'\}$ where $a_{mn}' = a_{mn}$ except that $a_{00}' = 0$. The calculations are as follows:

$$v = \sum_{m=0}^{N} a'_{mm} \Big/ \sum_{m=0}^{N} \sum_{n=0}^{N} a'_{mn} \tag{14}$$

$$X = \sum_{m=0}^{N} \left(\sum_{n=0}^{N} a'_{mn} * \sum_{n=0}^{N} a'_{nm} \right) \Big/ \left(\sum_{m=0}^{N} \sum_{n=0}^{N} a'_{mn} \right)^2 \tag{15}$$

$$\text{SEK} = e^{IOU_{e^{-1}}} \cdot (v - X) \Big/ (1 - X) \tag{16}$$

As part of SCD, the SS task and CD task are evaluated using MIoU and SeK.

5.3 Experimental Settings

The experiments are conducted on a Fujitsu Primergy RX2540 M1 server with CentOS 7, 128 GB RAM, 3.6 TB storage space and Intel Xeon E5–2630 processor @2.40 GHz. The proposed model is implemented using PyTorch library. The parameters set for the implementation includes: where batch size is assigned as 2, total epochs run is 50 and initial learning is set as 0.1. The gradient descent optimization method is SGD with Nesterov momentum.

5.4 Experimental Results

This section includes information about a series of tests are conducted to verify the efficacy of the proposed model for SCD and is compared with the several existing methods. The proposed model (ResCD-FCN) outperforms other models under observations because it uses attention like model (Resnet-34 backbone and ResCD block), which is optimised and can capture complex features, and because it uses FCN blocks, which are better at extracting features than standard methods using simple convolutional blocks. The UNET++ model, while functional, suffers from overfitting as it progresses down the model and may lose some important features, resulting in inaccurate results. The ResNet-GRU model is limited in its ability to retain information as it passes through multiple passes (timescales), which may result in inaccurate results. Although the ResNet-LSTM model outperforms the others under consideration, it suffers from the same disadvantage as ResNet-GRU, which may result in inaccurate results. The CNN-SCD model is limited in its ability to extract complex features and store temporal information in order to predict accurate results. The quantitative findings are shown in the following table (see Table 1). The results obtained by the proposed model are depicted below in a qualitative manner (see Fig. 5)

Table 1: Comparison between proposed model with literature methods for SCD

Methods	OA(%)	mIoU(%)	SeK(%)
UNET++	80.15	62.23	8.15
ResNET-GRU	79.97	60.38	7.75
ResNET-LSTM	82.84	63.76	13.57
CNN-SCD	79.86	61.65	7.65
ResCD-FCN	**84.04**	**67.67**	**14.12**

The labels specified in Fig. 5, T1, T2, GT1, GT2, S1, S2 corresponds to first year input satellite image, second year input satellite image, ground truth of first year satellite image, ground truth of second year satellite image, semantic change map of first year satellite image and semantic change map of second year satellite image respectively. These comparison tests indicate that the proposed model provides more precise results in SCD among the compared models. This model also has an advantage in embedding the semantic details which are particularly dominant.

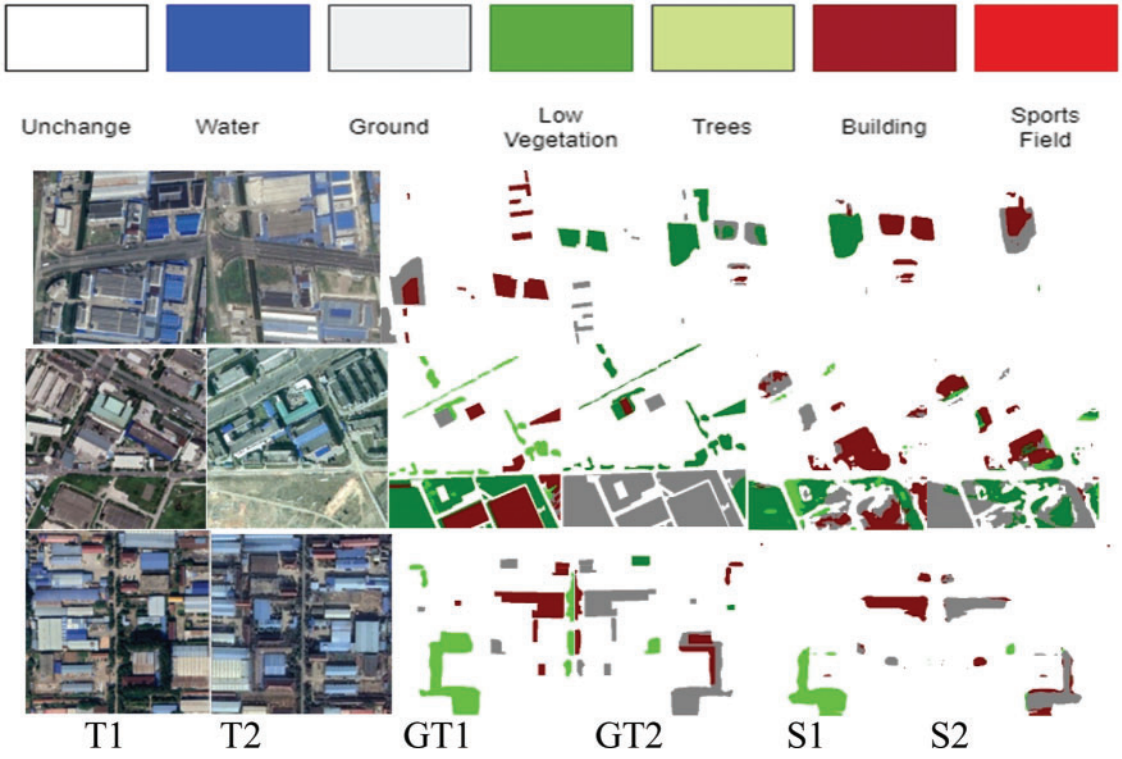

Figure 5: Results of the proposed model predicting semantic class changes

6 Conclusion

This research focuses to improve the semantic change detection task. The summarization of the various existing CNN architecture for SCD and identification of the drawbacks of these existing approaches are studied. The novel ResCD-FCN model, is proposed which merges the semantic features in CD block. Through the tests and the findings that emerge shows that, the proposed model outperforms other standard SCD architectures and State Of The Art (SOTA) methods and also obtain the highest accuracy when tested on SECOND dataset. Output of SS and CD tasks along with the loss function are merged and it helps to reutilize the semantic features in CD blocks which alleviates the accuracy of CD. Though this comparison between the proposed model and literature work architectures were carried out in CPU environment with batch-size of two, proposed model tends to be efficient. There is certain significant computational complexity to perform experiments; these limitations could be solved by working in a GPU environment with batch-size set as eight, by which proposed model will outperform in accuracy than all other related existing approaches works considered here.

One of the problems is to produce the time correlation between the semantic class changes especially in changed areas. Learning semantic class change conversion types may aid in the effective recognition of semantic classes. More connections between CDs and time partitions must be made to bring out these conversions, which is left for future work.

Acknowledgement: The authors would like to thank Department of Information Technology, Anna University, MIT campus for providing technical assistance and resources for this study.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] F. Bovolo and L. Bruzzone, "The time variable in data fusion: A change detection perspective," *IEEE Geoscience and Remote Sensing Magazine*, vol. 3, no. 3, pp. 8–26, 2015.

[2] F. Bovolo and L. Bruzzone, "A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 1, pp. 218–236, 2007.

[3] F. Bovolo, S. Marchesi and L. Bruzzone, "A framework for automatic and unsupervised detection of multiple changes in multitemporal images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 6, pp. 2196–2212, 2012.

[4] L. Bruzzone, R. Cossu and G. Vernazza, "Detection of land-cover transitions by combining multidate classifiers," *Pattern Recognition Letters*, vol. 25, no. 13, pp. 1491–1500, 2004.

[5] L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 3, pp. 1171–1182, 2000.

[6] L. Bruzzone, D. F. Prieto and S. B. Serpico, "A Neural-statistical approach to multitemporal and multisource remote-sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 3, pp. 1350–1359, 1999.

[7] L. Bruzzone and S. B. Serpico, "An iterative technique for the detection of land-cover transitions in multitemporal remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 35, no. 4, pp. 858–867, 1997.

[8] H. Kataoka, S. Shirakabe, Y. Miyashita, A. Nakamura, K. Iwata *et al.,* "Semantic change detection with hypermaps," arXiv preprint arXiv:1604.07513, vol. 2, no. 4, 2016.

[9] R. C. Daudt, B. Le Saux, A. Boulch and Y. Gousseau, "Multitask learning for large-scale semantic change detection," *Computer Vision and Image Understanding*, vol. 187, pp. 102783, 2019.

[10] J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, pp. 3431–3440, 2015.

[11] D. Peng, Y. Zhang and H. Guan, "End-to-end change detection for high resolution satellite images using improved unet++," *Remote Sensing*, vol. 11, no. 11, pp. 1382, 2019.

[12] K. Yang, G. -S. Xia, Z. Liu, B. Du, W. Yang *et al.,* "Asymmetric siamese networks for semantic change detection," arXiv preprint arXiv:2010.05687, 2020.

[13] R. C. Daudt, B. Le Saux and A. Boulch, "Fully convolutional siamese networks for change detection," in *2018 25th IEEE Int. Conf. on Image Processing (ICIP)*, Athens, Greece, IEEE, pp. 4063–4067, 2018.

[14] T. Suzuki, M. Minoguchi, R. Suzuki, A. Nakamura, K. Iwata *et al.,* "Semantic change detection," in *2018 15th Int. Conf. on Control, Automation, Robotics and Vision (ICARCV)*, Singapore, pp. 1785–1790, 2018. https://doi.org/10.1109/ICARCV.2018.8581264

[15] Y. Yuan and L. Lin, "Self-supervised pretraining of transformers for satellite image time series classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 474–487, 2021. https://doi.org/10.1109/JSTARS.2020.3036602

[16] K. Yang, G. -S. Xia, Z. Liu, B. Du, W. Yang *et al.,* "Asymmetric siamese networks for semantic change detection in aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2022, Art no. 5609818. https://doi.org/10.1109/TGRS.2021.3113912

[17] D. Peng, L. Bruzzone, Y. Zhang, H. Guan and P. He, "SCDNET: A novel convolutional network for semantic change detection in high resolution optical remote sensing imagery," *International Journal of Applied Earth Observation and Geoinformation*, vol. 103, pp. 102465, 2021, ISSN 1569–8432. https://doi.org/10.1016/j.jag.2021.102465

[18] S. Sun, L. Mu, L. Wang and P. Liu, "L-UNet: An LSTM network for remote sensing image change detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022, Art no. 8004505. https://doi.org/10.1109/LGRS.2020.3041530

[19] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng *et al.,* "A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022, Art no. 6506105. https://doi.org/10.1109/LGRS.2022.3143368

[20] Y. Hua, D. Marcos, L. Mou, X. X. Zhu and D. Tuia, "Semantic segmentation of remote sensing images with sparse annotations," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022, Art no. 8006305. https://doi.org/10.1109/LGRS.2021.3051053