RSCDNet: A Robust Deep Learning Architecture for Change Detection From Bi-Temporal High Resolution Remote Sensing Images

Deepanshi, Rahasya Barkur, Devishi Suresh, Shyam Lal, *Senior Member, IEEE*, C. Sudhakar Reddy, and P. G. Diwakar

Abstract*—Accurate change detection from high-resolution satellite and aerial images is of great significance in remote sensing for precise comprehension of Land cover (LC) variations. The current methods compromise with the spatial context; hence, they fail to detect and delineate small change areas and are unable to capture the difference between features of the bi-temporal images. This paper proposes Remote Sensing Change Detection Network (RSCDNet) - a robust end-to-end deep learning architecture for pixel-wise change detection from bi-temporal high-resolution remote-sensing (HRRS) images. The proposed RSCDNet model is based on an encoder-decoder framework integrated with the Modified Self-Attention (MSA) andthe Gated Linear Atrous Spatial Pyramid Pooling (GL-ASPP) blocks; both efficient mechanisms to regulate the field-of-view while finding the most suitable trade-off between accurate localization and context assimilation. The paper documents the design and development of the proposed RSCDNet model and compares its qualitative and quantitative results with state-of-the-art HRRS change detection architectures. The above mentioned novelties in the proposed architecture resulted in an F1-score of 98%, 98%, 88%, and 75% on the four publicly available HRRS datasets namely, Staza-Tisadob, Onera, CD-LEVIR, and WHU. In addition to the improvement in the performance metrics, the strategic connections in the proposed GL-ASPP and MSA units significantly reduce the prediction time per image (PTPI) and provide robustness against perturbations. Experimental results yield that the proposed RSCDNet model outperforms the most recent change detection benchmark models on all four HRRS datasets.

***Index Terms*—Remote sensing, deep learning, change detection, self attention, atrous spatial pyramid pooling.**

Manuscript received 22 November 2021; revised 6 March 2022, 2 July 2022, 18 September 2022, and 9 November 2022; accepted 13 December 2022. Date of publication 30 December 2022; date of current version 27 March 2023. This work was supported by RESPOND Scheme of Indian Space Research Organization(ISRO), Government of India under Grant ISRO/RES/4/683/19-20, December 30, 2019. *(Corresponding author: Shyam Lal.)*

Deepanshi, Rahasya Barkur, Devishi Suresh, and Shyam Lal are with the Department of Electronics and Communication Engineering, National Institute of Technology Karnataka Surathkal, Mangalore 575025, India (e-mail: deepanshi.asr.21@gmail.com; rahasyabarkur1999@gmail.com; devishisuresh@gmail.com; shyam.mtec@gmail.com).

C. Sudhakar Reddy is with the National Remote Sensing Centre, Indian Space Research Organisation, Balanagar, Hyderabad 500 037, India (e-mail: drsudhakarreddy@gmail.com).

P. G. Diwakar is with the Indian Space Research Organization (ISRO), Bangalore 560 094, India (e-mail: diwaa6@gmail.com).

This article has supplementary downloadable material available at https://doi.org/10.1109/TETCI.2022.3230941, provided by the authors.

Digital Object Identifier 10.1109/TETCI.2022.3230941

I. Introduction

The human civilization's settlement landscape and patterns are shifting rapidly due to the rise in human population, active urbanization, and various technologies. The task of change detection examines variations in natural phenomena or their state at regular intervals of time. The results thus obtained are processed to monitor urban expansion, environmental evaluation, plan a response to disasters, examine natural resources, and Land Use (LU) and Land Cover (LC) mapping. Keeping an accurate record, evaluating these differences, and rigorously analyzing this information is vital for better human development and to attain the Sustainable Development Goals (SDGs).

Change Detection (CD) systems attempt to assign a per-pixel binary label based on either bi-temporal (pair) or multi-temporal (sequence) co-registered images of a given area. The *changes* referred to in the problem statement can be due to changes in landscape, disappearing or construction of objects. A long withstanding challenge in the domain has been mis-classification of noise as semantic change. Atmospheric conditions, sensor calibration, geometric (viewpoint differences caused by camera zooming and rotation), and radiometric changes (illumination intensity variation, shadow, and seasonal changes), are the potential reason for noise in the collected sample. The broad latitude of object sizes arising from the method of data collection (aerial, satellite and drone images) and variability of change areas also add to the complexity of the problem. Although the distance between the change pixels and their noisy counterparts is narrow, we have used it to our advantage. By limiting the feature variations of unchanged pixels and highlighting the changed pixels, a change map is generated by the proposed architecture. Hence, feature comparison rather than image comparison forms the backbone of our proposed RSCDNet model to distinctly detect semantic changes and obtain satisfactory performance.

The initial exploration of the algorithms in change detection domain focused on separating the semantic and noisy changes. In general, traditional or unsupervised change detection algorithms can be classified into the following categories:

- *Image algebra practices* include image ratio, image regression, image difference, and change vector analysis [1], [2], among many others. In these algorithms, direct difference calculation on the bi-temporal images is performed.

- *Image transformation algorithms* like Principal Component Analysis (PCA) [3], Multivariate Alteration Detection (MAD) [4], and Independent Component Analysis [5] extract useful features from bi-temporal images by altering and merging their feature bands.
- *Classification methods* include compound classification and post-classification; both methods for obtaining land-use categories [6], [7].
- Markov Random Fields (MRFs), wavelets, and local gradual descent Algorithms [8], [9], fall into the category of advanced models.

Programs like Copernicus and Landsat have made massive volumes of Earth observation imagery available. World-View, QuickBird, DeepGlobal, GF1, ZY-3, Sentinel, GF2, and several other satellite sensors' data can now be used to build advanced remote sensing technologies. The boom of deep learning coupled with massive amount of aerial data lead to development of CD methods based on the same which can be classified as

- Feature-Based: These are designed to execute feature engineering independently. These algorithms study the data to search and correlate the features to facilitate more active learning without being told [10].
- Patch-Based: Pixel patches are created using either raw or difference images which are then supplied to a deep learning model to determine the center pixel's change association [11].
- Image-Based: The principle of segmentation underpins this approach. Segmentation produces results from other images using end-to-end training thus minimizing the impact of pixel patches as much as possible [12].

The proposed RSCDNet model, incorporating GL-ASPP and MSA blocks for context assimilation and accurate localization falls in the Image-Based category.

The following is a summary of the rest of the paper. Relevant work is presented in Section II. The development and design of the proposed deep learning framework, RSCDNet is discussed in Section III. The specifics of the training and implementation, as well as the experimental results, are covered in Sections IV and V. The conclusion of this paper is delineated in Section VI.

II. Related Work

The interpretation of changes from bi-temporal images can be done through two primary standards: unsupervised and supervised. The former generates binary maps where the user's influence is minimal, while the latter requires a labeled set of examples and is more suited for real-world situations.

Initial attempts at identifying changes from bi-temporal images were made in the early 2000 s using Markov Random fields [2] to exploit the inter-pixel class dependencies. With the development of image processing methods, the architectures in [3], [13] generated change maps from the difference features of bi-temporal images through linear transformation and clustering. However, these techniques necessitated fine-tuning of multiple parameters for different datasets. This issue was subsequently resolved in [14] which introduced unsupervised Bayesian frameworks that binarised the difference image to generate change maps. R Touati et al. in [15] further developed the Bayesian framework by incorporating the Markov mixture model and Maximum a posteriori (MAP) solution with the stochastic optimization procedure. However, due to the grayscale conversion of the input images, the algorithm suffers from a lack of chrominance information, and addition of noise owing to bilinear interpolation. The failure of the above outlined pixel-based techniques to collect all the contextual information prompted the development of object-oriented methodologies [16], [17]. For each bi-temporal image, the authors in [16] obtained a sparse description of the object features using five statistical key points to mitigate the negative effects of outlier pixels which were subsequently evaluated to construct change maps. Notwithstanding the improvised performance, the model falls short in the detection of multiscale objects. A shift from statistical methods to invariant image modality is observed in [18] and [19]. The architecture in [18] presented an independent concentric circular invariant convolution which projects the first bi-temporal image onto the imaging modality of the second image. The model, however, fails to identify the change when the variability between the bi-temporal images is high. In [19], imaging modality-invariant operator forms the basis of the algorithm. The variations in every structural area in the two bi-temporal images were identified in terms of the high-frequency pattern.

The results from the unsupervised methods fail to capture the more delicate patterns to draw a change map from bi-temporal images. The high noisy changes due to varied illumination and viewpoints lead to low precision in the computed change map. The availability of high computational power and an enormous amount of data from various satellite missions such as Hi-UCD [20] has paved the way for supervised methods producing more reliable results.

The work in [21] was one of the first endeavors to use spectral data to train a multi-layer perceptron network. The advent of Convolutional Neural Networks (CNNs) transformed the given input to the intended outcome through a series of processing steps, resulting in a hierarchy of feature maps as seen in U-Net [12]. A comprehensive overview of current breakthroughs in deep learning-based change detection techniques can be found in [22]. The state-of-the-art architectures can be divided into two categories: early concatenation and feature comparison. [23] is one of the earliest works in the early concatenation sub-category. The algorithm divides the bi-temporal images into patches of size $(15 \times 15 \times C)$, which are then concatenated and processed through multiple convolutional and fully connected layers to obtain a change map. However, the absence of spatial awareness resulted in erroneous change boundaries as well as a lack of attention to detail. The research in [24] tackled this problem by introducing LUNet, an end-to-end spatiotemporal network that integrates LSTM-Conv layers. The inability to extract deep and complicated features was overcome as a result of this advancement. This obstacle of overlooking finer details was partially solved in Peng et al. [25]. The authors proposed UNet++ [25], a variant of the U-Net [12] incorporating convolution blocks in skip connections between the encoder and decoder which helps alleviate the semantic gap. However, architecture has a limited field of view and neglects global relations. The AGCDetNet [26] architecture integrates an extensive dilated convolution and spatial attention unit to capture complete context for each pixel. Despite perceived improvements, the model is computationally expensive, and direct interpolation of the squeezed features results in loss of information. The above-mentioned early concatenation algorithms are image-based; hence they are prone to misclassifying noisy changes as semantic ones. This is where feature comparison models such as Siamese [11], DSMSCN [27] and UCDNet [28] come into the picture with their ability to correlate and connect object-based characteristics from the bi-temporal images. In [11], the siamese architecture independently derives features through downsampling the two input images and eventually uses the pixel-wise Euclidean distance to produce a change map. However, the model overlooks the contextual relationship of the adjacent pixels while generating the thresholded change map. The above-mentioned architecture was improved by extracting characteristics from bi-temporal images using pre-trained Sub VGG16 weights [29]. Nevertheless, since these weights were engineered for image scene classification, they need to be fine-tuned for the change detection task. Mengya Zhang et al. [30] further built upon the siamese architecture by optimising it with the modified triplet loss in order to improve intraclass inseparability and inter-class separability. The model yields more reliable results, direct interpolation of the extracted feature maps to the size of the input image resulted in the addition of noise and loss of object information. Another attempt to modify the siamese model is [27] through the integration of a decoder unit for upsampling instead of interpolation. Furthermore, the study utilizes skip connections at each layer to take advantage of not only sophisticated traits but also variances in lower-level features.

The literature presented above sheds light on the lack of harmonious balance between context assimilation and accurate localization. Furthermore, most of the preceding architectures incorporate premature concatenation which erroneously classifies noisy changes as semantic ones. This propelled us to incorporate feature extraction followed by comparison since noise in the bi-temporal images exhibits similar distribution. Here we present RSCDNet, an end-to-end trainable model with Gated Linear Atrous Spatial Pyramid Pooling (GL-ASPP) and Modified Self-Attention (MSA) blocks to better capture the contextual information.

Novelties of proposed RSCDNet architecture are as follows:

- The two branches of the encoder unit share weights to derive similar complex traits from the bi-temporal images. The extracted characteristics are then subtracted at the bottleneck for feature comparison. This mechanism also helps the proposed model to be more robust to the synthetic perturbations.
- The proposed model effectively utilizes a Modified Self-Attention (MSA) mechanism at the bottleneck to incorporate the spatial dependencies existing in the obtained features. A gate is integrated into one of the feature spaces in the MSA block to highlight the salient channels.
- The processed features are subsequently passed through a newly introduced 6-Level Gated Linear Atrous Spatial Pyramid Pooling (GL-ASPP) block to capture a larger field of view. The addition of the gated linear unit after concatenation of the dilated features helps to suppress irrelevant channel information.

III. Proposed Architecture

The detailed functionalities of the Gated Linear Atrous Spatial Pyramid Pooling (GL-ASPP) block and Modified Self Attention (MSA) module which are included in the proposed RSCDNet Model are presented in the Sections III-A, III-B.

A. Gated Linear - Atrous Spatial Pyramid Pooling

Dilated convolution enables the model to expand the field of view of kernels without compromising on computational cost thus allowing it to integrate additional contextual information. Atrous Spatial Pyramid Pooling (ASPP) block [31] employs multi-rate dilated convolution in parallel with spatial squeeze function to detect objects of varying sizes.Using this as a foundation, we designed a Gated Linear Atrous Spatial Pyramid Pooling (GL-ASPP) assembly mounted with a channel-wise descriptor and gated module to remove redundant channel information from multi-scale features. In contrast to the vanilla ASPP unit [31], we have limited the role of the presented unit to widening the field of view rather than using it in combination with other techniques to acquire complete context. This helps the block to focus on selective neighbor context assimilation. The newly introduced GL-ASPP block also takes into consideration the channel dependencies to filter out peculiar information flow from the encoder to the decoder.

The GL-ASPP block as shown in Fig. 1 comprises of six parallel dilated convolutions with varying dilation rates and a channel descriptor gate. The channel descriptor gate calculates attention weights with squeeze operation and computes channel-wise product. The refined and highlighted multi-scale characteristics from each branch are then aggregated and fed into a Gated Linear Unit. The input is parallelly processed through tanh and sigmoid activation branches which suppress irrelevant contextual information. This output is fused with the input to the GL-ASPP to get a hybrid selection of advanced contextual features from a larger neighborhood.

Let $X^i \in R^{H\times W\times C}$ (i=1,...6) be the simultaneously obtained output from the six branched dilated convolution. Each feature vector, X^i, is multiplied with the channel attention weights computed using the global average pooling denoted by ζ in (1).

$$GA_i = X^i \otimes \zeta(X^i) \tag{1}$$

The output from each branch is then concatenated to generate a vector $GA \in R^{H\times W\times 6C}$ which acts as an input to the Gated Linear Unit. Let τ and σ denote the tanh and sigmoid activations. Then the output of the linear unit, GLU, can be expressed as in (2).

$$GLU = \tau(GA) \otimes \sigma(GA) \tag{2}$$

The GLU output is passed through a 1×1 convolution, which produces $X_t \in R^{H\times W\times C}$. Finally, a skip connection from the GL-ASPP input is added to promote convergence.

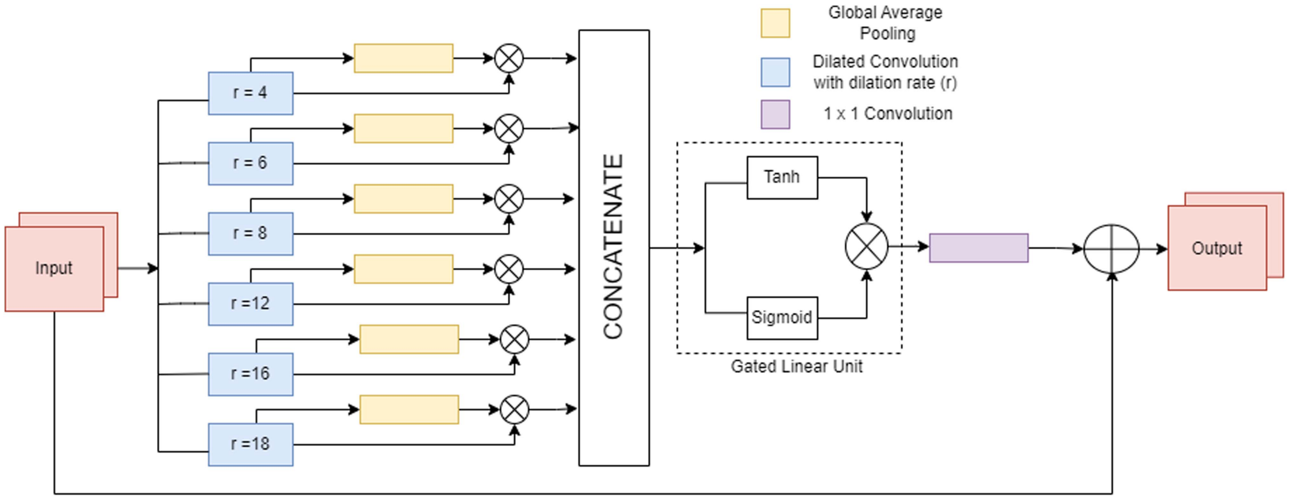

Fig. 1. Proposed sub module: Gated Linear ASPP block (GL-ASPP).

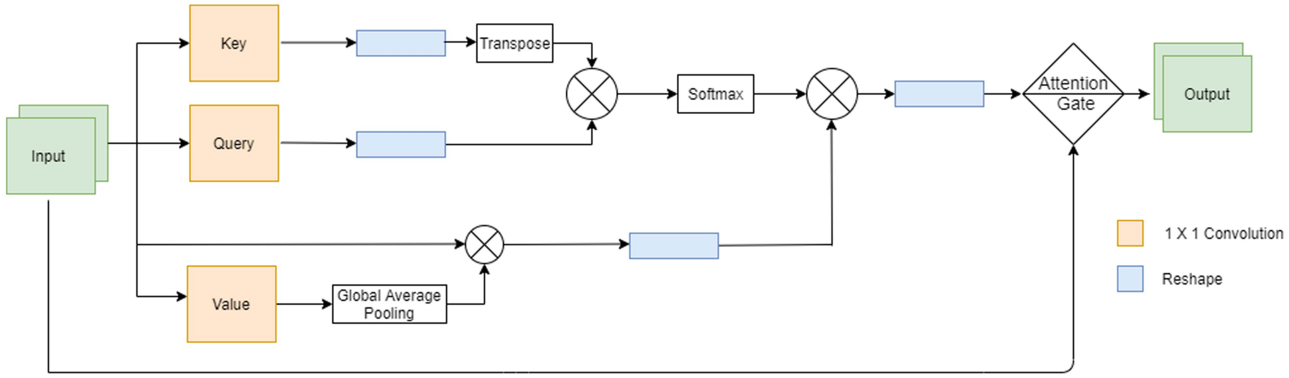

Fig. 2. Proposed sub module: Modified Self Attention block (MSA).

B. Modified Self-Attention

Self-Attention, also known as intra-attention, is a mechanism to compute global dependencies for each pixel. In this mechanism [32], all pairing co-variances between the pixels are calculated which is then used to enhance or weaken its value according to its similarity with other pixels in the feature map. The receptive field of self-attention is the complete feature map as opposed to the convolution operator. The Modified Self Attention (MSA) block is systematically designed to combine the functionality of both the backbone self-attention unit and channel attention operation. This operator exploits the spatial-channel interdependence of the input feature vector. The computed channel-self attention characteristics are then transmitted via an attention gate to filter extraneous information and emphasize the crucial ones contrary to the vanilla self-attention module [32]. By overlaying the learned attention map over the deep features, the attention gate enhances differentiation between changed objects and the background.

The MSA block shown in Fig. 2 generates a weighted aggregate of the neighborhood vectors using three concurrent branches. To generate an intermediate value feature, $V_i \in R^{1\times 1\times C}$, a 1×1 convolution is performed on the input vector, $X_i \in R^{H\times W\times C}$ followed by ReLU activation and global average pooling. The final feature space, v, is obtained by fusing the computed channel weights with the input vector and it is given in (3).

$$v = V_i \otimes X_i \tag{3}$$

A similar systematic approach is adopted to compute the vectors in the other two branches. The two feature spaces, query (q) and key (k) are obtained by a 1×1 convolution with ReLU activation and reshaped into $R^{N\times C_f}$, where $N = H \times W$ and $C_f = N_c/F$ (N_c and F are 256 and 8 respectively). q and k are multiplied (element-wise) to generate pixel-wise correlation or attention weights (γ). The outcome is then normalized through the softmax operation and is given in (4).

$$\gamma = \iota[q^T \otimes k] \tag{4}$$

where ι, T and $\otimes$ indicate softmax, transpose, and element-wise matrix multiplication operation respectively.

The intermediate vector γ, and v from the above branches, are multiplied and reshaped to produce a multi-layered, highly sensitive feature cluster, $\varphi \in R^{H\times W\times C}$ which needs to be further refined to suppress redundant features and highlight the salient ones. A very practical and structured solution is the addition of an attention skip connection [33]. The multi-step delineation of the MSA block culminates in the processing of φ and the input feature vector. The penultimate operation of the MSA block is carried out inside the attention gate. Let a, b be the intermediate features obtained after 1×1 convolution applied on the input feature vector X_i and $\varphi \in R^{H\times W\times C}$ respectively. The fused output of a and b is scaled with the ReLU operator and then passed through a1×1 convolution with sigmoid activation, to generate weights, ψ blue$\in R^{H\times W\times 1}$.

$$\Upsilon = \psi \otimes X_i \tag{5}$$

The output, Υ in (5) is an aggregation of highly refined features. The MSA block thus serves its purpose as a global pixel dependency calculator.

C. Explanation and Mathematical Representation of Proposed Architecture and Its Intermediate Stages

The aim of the research discussed in this paper is to find binary changes in bi-temporal images. The final proposed architecture results from adopting a systematized approach of experimentation with different techniques and selective assimilation of those that yield a significant improvement over the existing models. Each intermediate stage explained in the following sections delineates the thought process and detailed design analysis of each framework which finally builds up to the proposed RSCDNet model.

1) Intermediate Stage I: Intermediate stage I is based on early fusion [23] and is modified to obtain a detailed description of the change map. The contracting path is a VGG-16 based network consisting of two 3×3 recurrent convolutions (padded) with a rectified linear unit (ReLU) activation, a 0.2 rate dropout regularisation to palliate over-fitting and a 2×2 max pooling operation for down-sampling. Thus after each contraction stage, the spatial complexity is lowered while the feature information is intensified. The obtained feature vectors at the end of the contracting paths are concatenated for comparison at the bottleneck. Each step in the expansive path consists of upsampling the characteristics which doubles the spatial dimension and halves the number of features. It is followed by a repeated 3×3 convolution, ReLU activation, and a dropout layer. The processed feature maps are passed through a final 1×1 convolution with sigmoid activation to generate the binary change map.

2) Intermediate Stage II: The proposed architecture's second intermediate stage adds skip links between the encoder and decoder units. A skip connection, as the name suggests, tends to skip some of the network layers. It provides an alternative path for gradients to disseminate during back-propagation and aids in resolving the problem of vanishing gradients thereby making it ideal for model convergence. These connections pass the information to the layers at the end of the architecture, making it easier to classify the minute details.

Every skip connection in an n-step architecture merely concatenates all features at layer j with those at layer n-j. The features of both encoders are added and then fused with the features of the ensuing decoder. The fused output, $x[j]$ can be expressed as follows:

$$x[j] = F(x[n-j] \oplus (x_1[j] + x_2[j])) \tag{6}$$

In (6), $x_1[j]$ and $x_2[j]$ are the j^{th} step outputs for each encoder and $\oplus$ is concatenation of the sum with the $n-j$ step decoder output.

3) Intermediate Stage III: In the third intermediate stage of the proposed RSCDNet architecture, we introduce the vanilla ASPP block [31] at the bottleneck. As the task of change detection demands the perception of objects of differing sizes, it is imperative to introduce the ASPP block to increase the field of view for each pixel. The input feature vector is operated upon by dilation convolutions and eventually merged.

The ASPP block entails 4 sub-blocks performing atrous convolutions of discrete rates followed by a dropout layer to reduce overfitting. The resultant features from each sub-block are added and subsequently up-sampled in the decoder block.

$$T = \sum_r Y_r \qquad \forall r \in [r_1, r_2, r_3, \ldots., r_n] \tag{7}$$

where Y$_r$ is the output from (1) and r depicts the rate used for dilated convolution. We experimented with different dilation rate series to get the fine trade-off between the small and large fields of view. The two rate series which performed modestly were 4-8-12-16 and 6-12-18-24, the results from which are compared in the next section.

4) Proposed RSCDNet Model (Final Stage): A thorough analysis of the intermediate stages described above shows that the task of change detection can be better performed by amplifying contextual perspectives for higher performance metrics. In the final proposed RSCDNet architecture n Fig. 3, Modified Self-Attention (MSA) and a Gated Linear Atrous Spatial Pyramid (GL-ASPP) Pooling unit are embedded at the bottleneck. Instrumental errors in data measurements are unavoidable due to the precision limits in construction and also due to wearing over time. This combined with changed weather patterns introduce noise to the data we wish to study. Extracting similar features from the image data for the purpose of change detection and their subsequent comparative analysis in the deeper layers of the model reduces this noise since this noise in both the images will have similar distributions.

The input is transformed to a latent representation by the dual encoder branches, and the decoder reconstructs a binary change map from it. We enabled a weight-sharing mechanism in the dual encoder network to extract similar features, which are subsequently subtracted. The subtraction allows for the retention of essential characteristics pertaining to changed areas while suppressing the redundant information. This process reduces parameters and makes the model computationally efficient. The processed features are then passed through an extensive MSA and GL-ASPP block. In comparison to the existing architectures, we introduce a bifurcation into global context assimilation and increased field of view for each pixel. This twin process performs complementary functions; the first one is concerned with change areas in a global context, while the other caters to neighborhood pixels. The modified self-attention unit acts as a

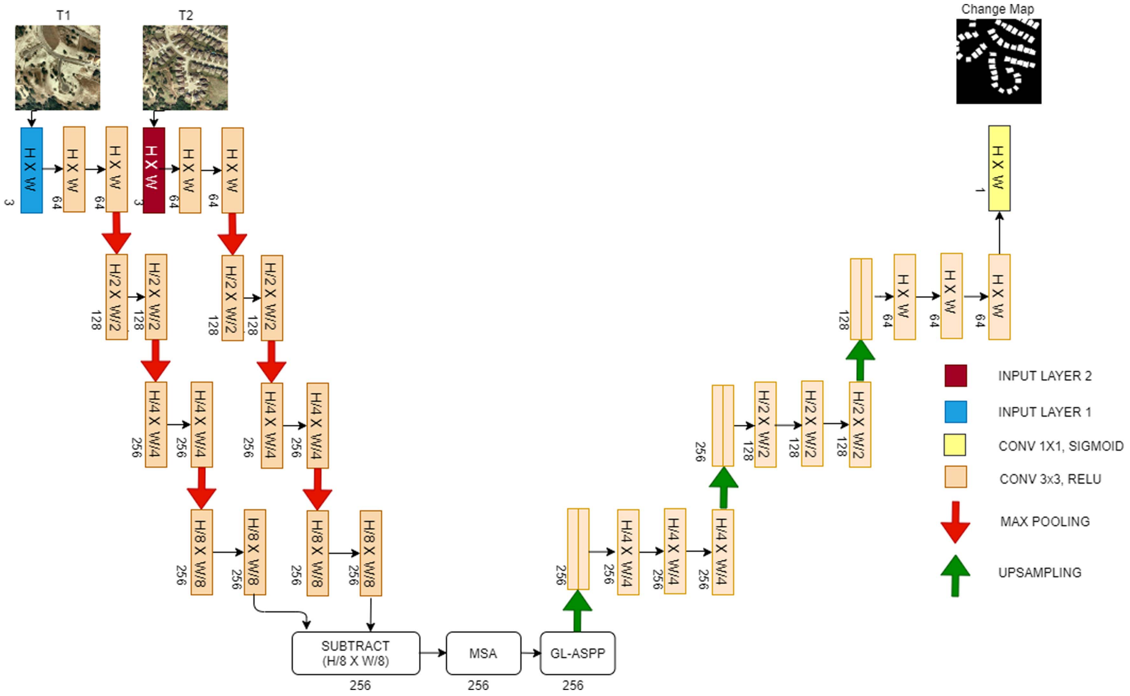

Fig. 3. Block diagram of the proposed RSCDNet model.

context calculator for each pixel and increases or decreases the vector accordingly, while the GL-ASPP block works to improve the traced boundaries of the changed areas.

The features from the two contracting encoder paths are subtracted and fed into the MSA block. If X_1 and X_2 are the outputs from the contracting paths for the first and second images respectively, then subtracted output, X can be expressed as in (8). This is then passed through the Modified Self-Attention block ((3), 4 and 5).

$$X = X_1 - X_2 \tag{8}$$

$$L = \alpha(X) \tag{9}$$

In (9), α is the modified self-attention operation. After observing the results of both the rate series from the intermediate stage III model, the GL-ASPP block is constructed with varying rates to precisely capture the local context. The analysis of our extensive experimentation justified the adoption of the rate series 4-6-8-12-16-18 as the most suitable for the task of change detection. The intermediate result, before applying GLU, can be expressed as in (10).

$$T = \sum_r Y_r \qquad \forall r \in [4, 6, 8, 12, 16, 18] \tag{10}$$

where Y_r is the output from the (1) and r depicts the sampling rate used for dilated convolution. The output T is then passed through the GLU operator of (2) to nullify superfluous characteristics. Finally, a skip connection from the input to the GL-ASPP output is added and it is given in (11)

$$O = GLU(T) + L \tag{11}$$

Thus, with the incorporation of the MSA and GL-ASPP block, a trade-off between enlarged and diminished fields of view with minimal loss of information is achieved.

IV. Training and Implementation Details

A. Dataset

To develop an effective change detection architecture that can recognize all types of changes from reconstruction/demolition of buildings to natural changes, the proposed architecture, its intermediate stages, state-of-the-art architectures, and the existing segmentation models are all trained and evaluated on four different datasets.

1) SZTAKI Air Change Benchmark Set (Dataset I) [34], [35]: This is a publicly accessible benchmark dataset that contains 13 aerial RGB image pairs with a resolution of 1.5 m/pixel and a shape of 952 × 640 pixels, as well as binary change masks that were used for [34], [35] publication assessment. The input images were taken over a period of 23 years. For change mask labeling, the following differences in the bi-temporal images are labeled as changed regions: (a) newly constructed regions (b) building operations (c) planting of large groups of trees (d) fresh plow-land (e) groundwork before building over. Each image in the provided dataset is padded to reshape it to 1024 × 640 pixels and augmented with y-axis flips, 90 degrees rotations, and modifying the brightness of the images by a random factor. There are 48 training images and 12 testing images in total.

2) Onera Satellite Change Detection (Dataset II) [23]: The Onera Satellite Change Detection (OSCD) dataset was acquired from the Sentinel-2 satellites. Twenty-four areas with an average size of 600 × 600 pixels and a resolution of 10 meters

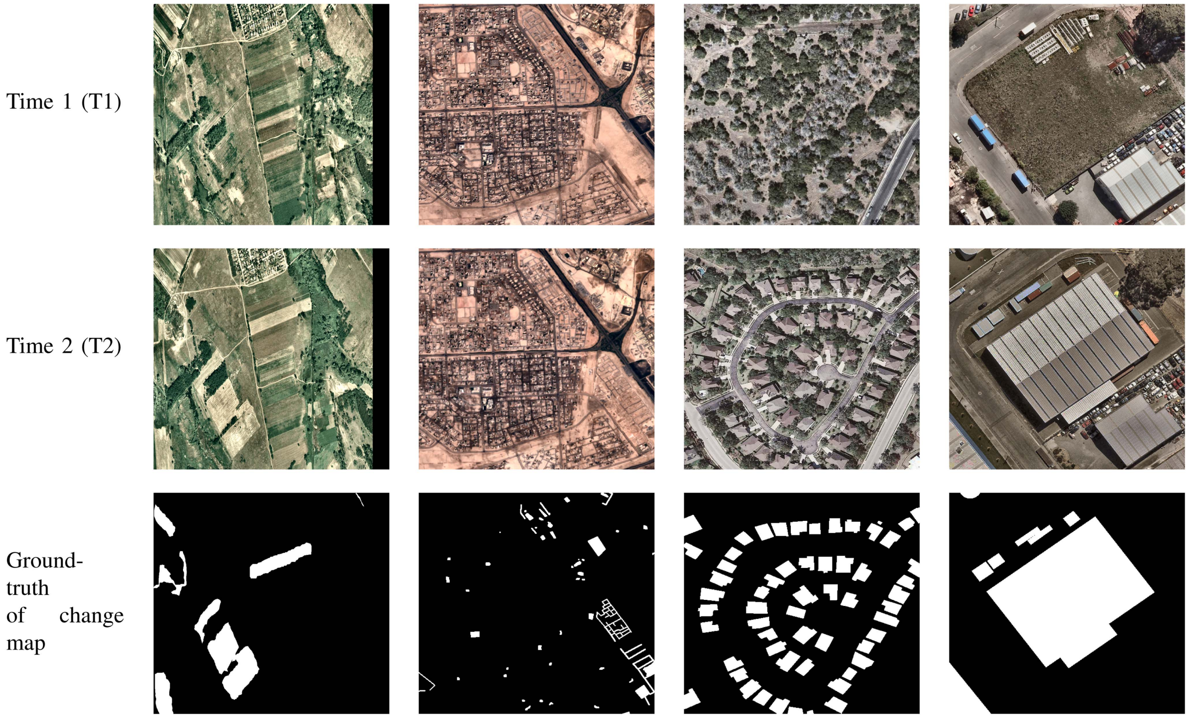

Fig. 4. Test images of the datasets Col I: Dataset I (23 years of time difference), Col II: Dataset II (between 2015 and 2018), Col III: Dataset III (between 2002 and 2018), and Col IV: Dataset IV (between 2012 and 2016)

were selected from across the world, with differing levels of urbanization and evident urban changes like buildings, roads, etc. All of the images were cropped according to the preferred geographical coordinates, yielding 13 band image pairs. Our work is focused on only three band images. We considered only the RGB bands and each image in the dataset was split evenly into patches with a spatial resolution of 512×512 and a 64-pixel overlap between neighboring patches. There were 200 images for training, 30 for validation, and 57 for testing. The focus of this dataset is on urban changes.

3) LEVIR - CD (Dataset III) [36]: This is a public dataset available through the LEarning, VIsion, and Remote sensing laboratory mentioned in paper [36]. It's composed of 637 Google Earth (GE) pairs of image patches with a resolution of 0.5 m/pixel and a size of 1024×1024 pixels. Significant land-use changes, particularly the construction of man-made objects, roads, buildings. etc can be seen in these bi-temporal images which span 5 to 14 years. Given the higher spatial complexity of these images, we divided each image into four 512×512 patches without overlapping of neighboring patches. 1780 training images, 256 for validation, and 512 for testing were generated from the dataset. This dataset covered seasonal and illumination variations, which contributed to the development and implementation of effective change detection methods that can reduce the influence of unnecessary changes over real changes.

4) WHU Building Dataset (Dataset IV) [37]: The reference change masks and a pair of co-registered aerial images (TA-2011 and TA-2016) with a combined size of $15,354 \times 32,507$ pixels. The research is being conducted in Christchurch, New Zealand, which was affected by an earthquake in 2011. The study region encompasses a significant amount of new construction. These pictures have a ground sampling distance of 0.2 m/pixel. For model training and evaluation, we cropped the original photos into smaller image tiles with a size of 512×512 pixels. For training purposes, patches with the unchanged class were undersampled, creating 554 samples while 660 samples for testing purposes were generated.

One set of test case of bi-temporal high-resolution remote sensing images from every dataset is given in Fig. 4 and three sets of test cases of bi-temporal high-resolution remote sensing images are documented in the supplementary file.

B. Loss Function

The weighted binary cross-entropy loss (L_{wce}) was used in the change detection process. The performance of a model whose output is a probability value between 0 and 1 is assessed by cross-entropy loss [38], otherwise termed as log loss. As the predicted probability differs from the actual label, cross-entropy loss increases. The binary crossentropy loss can be mathematically represented as in (12)

$$L(t, \hat{t}) = -\left[\frac{1}{N}\sum_{i=1}^{N} t_i \times log(\hat{t_i}) + (1 - t_i) \times log(1 - \hat{t_i})\right] \tag{12}$$

The distribution of changed and unchanged pixels is heavily biased for the change detection task. For instance, the distribution of unchanged and changed pixels in dataset III is 95% and 5%, respectively. As a result, using only binary cross-entropy as the

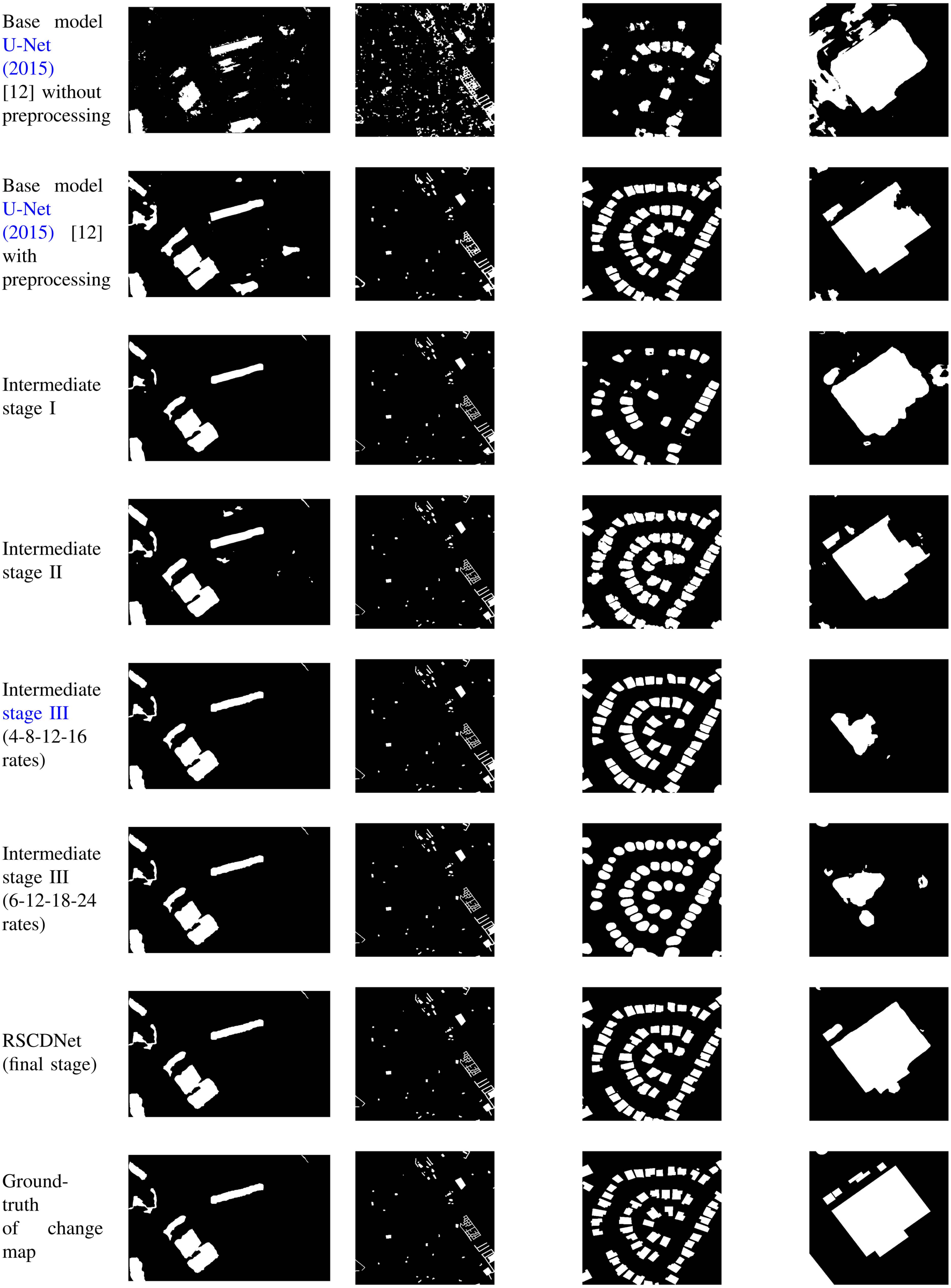
Base model U-Net (2015) [12] without preprocessing
Base model U-Net (2015) [12] with preprocessing
Intermediate stage I
Intermediate stage II
Intermediate stage III (4-8-12-16 rates)
Intermediate stage III (6-12-18-24 rates)
RSCDNet (final stage)
Ground-truth of change map

Fig. 5. Predicted change map of proposed RSCDNet and its intermediate stages with base model U-Net (2015) [12]. Col I: Dataset I, Col II: Dataset II, Col III: Dataset III, Col IV: Dataset IV.

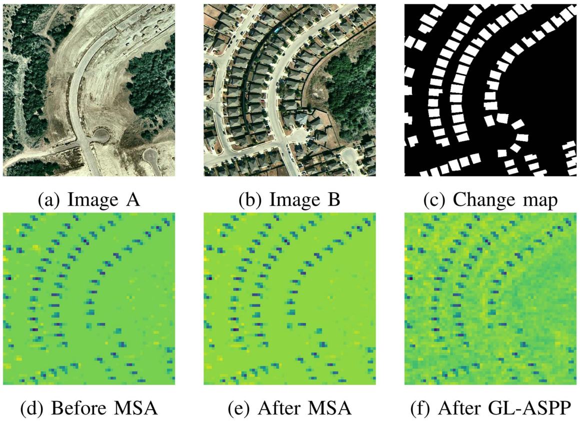

Fig. 6. Visualization of the intermediate stages in RSCDNet.

loss function creates a class imbalance problem. To solve this issue, a simple weighted balance strategy has been used, which is defined as:

$$L = -\left[\sum \alpha(t_j) \times \log(\hat{t_j}) + \eta(1 - t_j) \times \log(1 - \hat{t_j})\right] \tag{13}$$

In (13), α and η are the weights to adjust the imbalance between the classes and penalize the loss function more for false-positive predictions. Extensive experimentation led to the selection of α as 2.5 and η as 0.5.

C. Preprocessing

The bi-temporal images were passed through a pre-processing unit to account for varying radiometric changes such as illumination intensity, shadow, and seasonality. A histogram matching and contrast enhancement operation were performed in this section. Histogram matching is the means of transforming the histogram of a time series, image, or higher-dimensional scalar data such that it matches the histogram of another reference. Contrast Limited Adaptive Histogram Equalization (CLAHE) [39] was applied to each of the images after histogram matching. Instead of global equalization, CLAHE boosts the image's contrast locally. This method of local equalization works to prevent over-amplification of noise in the image's near-constant regions.

D. Training Setup

A 32-bit operating system with an 8 GB VRAM [NVIDIA Quadro P4000 GPU] was used for the training of all models on Dataset II and Dataset III. Google Colab [NVIDIA Tesla K80 GPU] with 12 GB VRAM is used to train models on Dataset I as it required a lot more RAM in comparison to other datasets. We adopted TensorFlow 2.0 with the Keras API framework for training and evaluating all the models. The Adam optimizer was used to train all models with a learning rate of 0.0001, $beta_1$ of 0.9, $beta_2$ of 0.999 and epsilon of 1×10^{-7}. All models show convergence for Dataset I, II, III, and IV after 100, 150, 50, and 100 epochs respectively. The python source code of the proposed model will be available at[1]

V. Results and Discussion

A. Quality Metrics

Performance metrics such as F1 score, precision, Recall, Jaccard score, Kappa coefficient, Overall Accuracy (OA), True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), and False Negative Rate (FNR) [40], [41] are used to quantify the results of binary change map generated from processing bi-temporal high-resolution remote sensing images.

B. Ablation Study

In this section, we undertake the task of comparing the effects of incorporation of the meticulously designed MSA and GL-ASPP block to the proposed RSCDNet model based on qualitative results and performance metrics. Tables I, II, III and IV present a comparative analysis of the quality metrics on four different datasets. Fig. 5 presents a qualitative insight on the results obtained from the proposed RSCDNet model with its intermediate stages for one of the test cases. Observing Fig. 5 it can be concluded that the intermediate stage I correctly identifies large change areas but fails to capture the smaller counterparts. With the inclusion of skip connections in intermediate stage II, the model overcompensates and falls short of correctly differentiating between semantic and noisy changes, as can be seen from the predicted change maps of Dataset I (Fig. 5). The vanilla ASPP network in the intermediate stage III, leads to a raise of 12% in the Kappa coefficient of Dataset I along with an increase in the identified change areas in the predictions of stage III. Furthermore, using the rate series 6-12-18-24 in the dilation convolution branches, all the change areas are predicted; however, the boundaries of the changed areas are irregular as seen in Dataset III's predicted change maps. While the amount of contact between the boundaries is reduced in intermediate stage III with the 4-8-12-16 series, a significant portion of the change area is mis-classified, resulting in a low recall score. We combined the two series to create the 4-6-8-12-16-18 dilation rate sequence for the GL-ASPP network in the proposed RSCDNet model to get the best of both worlds. The performance of the proposed RSCDNet is further enhanced through the integration of MSA and GL-ASPP blocks which ameliorates boundary detection of changed areas by capturing the global context; the predicted change maps of Dataset IV are a testament to this improvement. Unlike the intermediate stages, the proposed RSCDNet has shared weights in encoders, the processed feature maps from which are then subtracted and not concatenated. Apart from reducing the number of parameters, another added advantage of shared encoders, is the removal of noise. The calibration or sensor noise in the bi-temporal images will have similar distribution across the range, which is hence truncated by subtracting the encoder feature vectors and not the images directly.

[1] [Online]. Available: https://github.com/deepanshi-s/RSCDNet

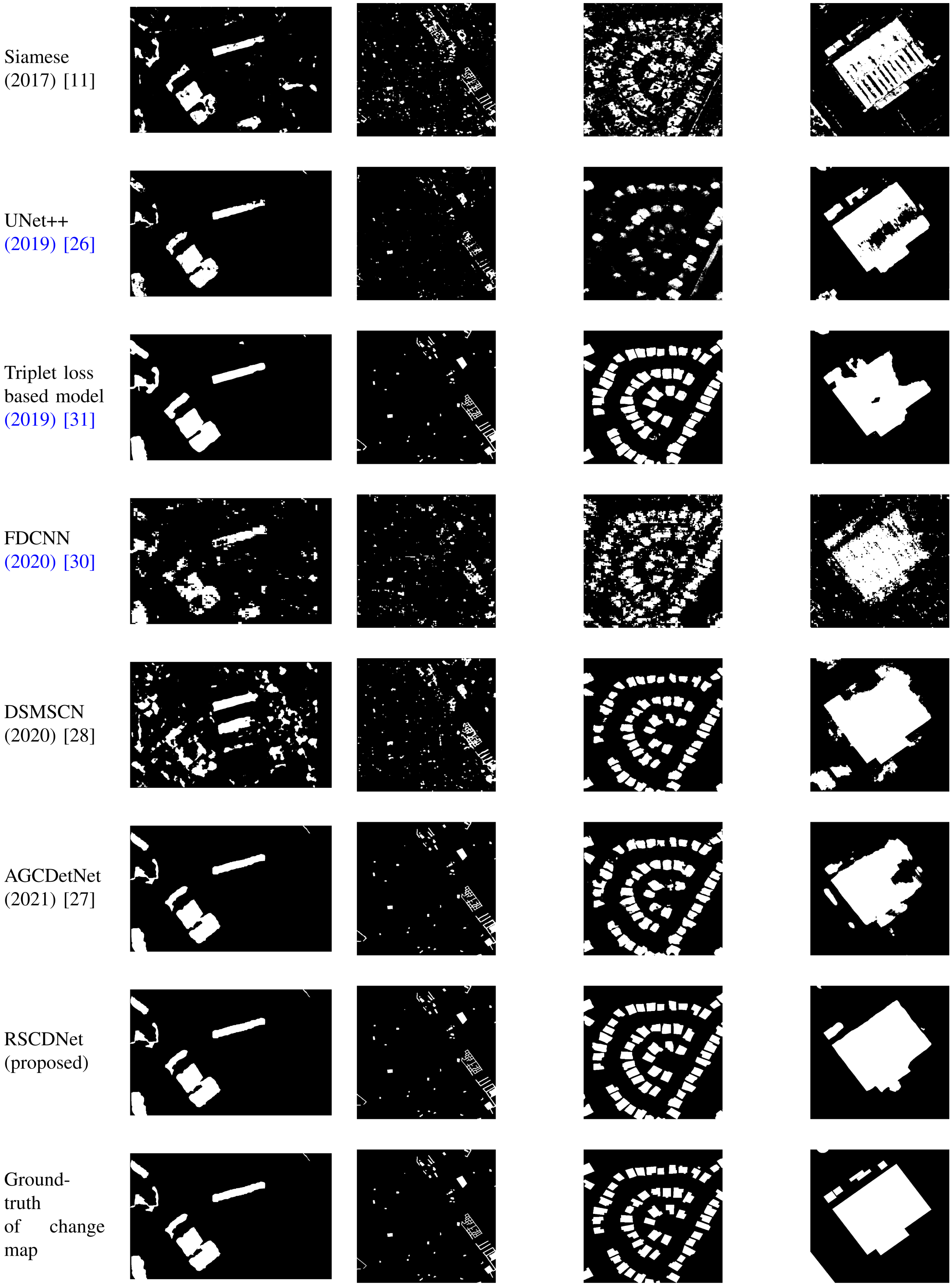

Fig. 7. Predicted change map of proposed RSCDNet and other existant deep learning models on the test images. Col I: Dataset I, Col II: Dataset II, Col III: Dataset III, Col IV: Dataset IV.

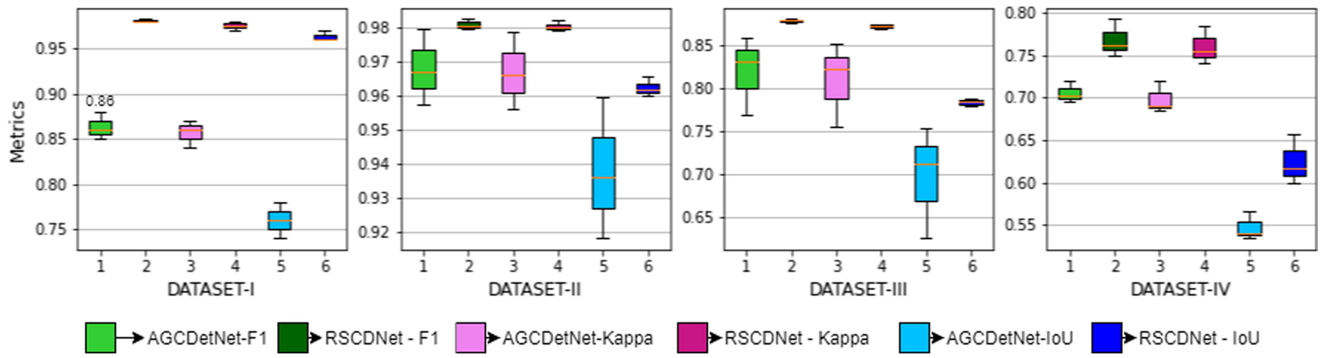

Fig. 8. Box plot of RSCDNet and AGCDetNet.

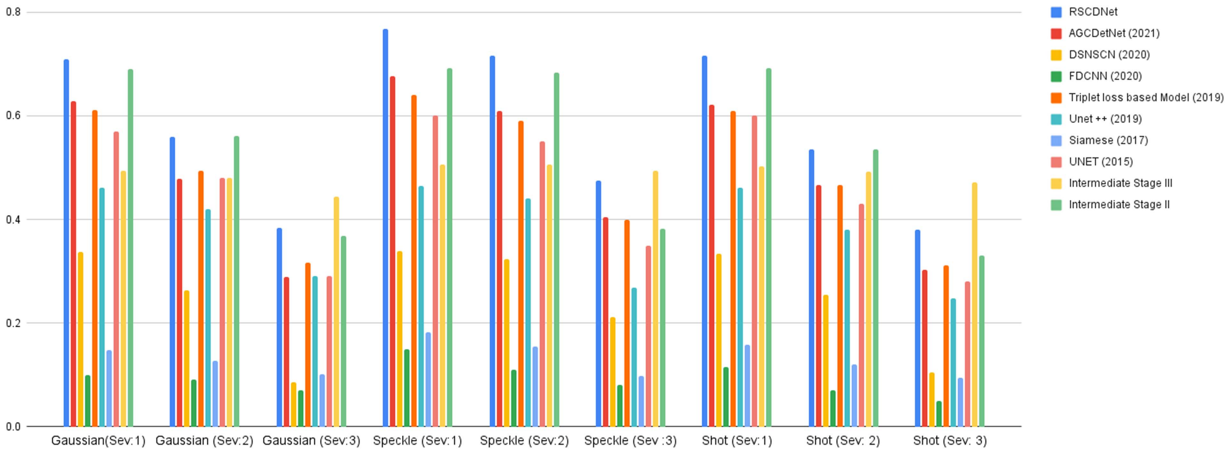

Fig. 9. Performance (F1-score) degradation comparison of the state-of-art models with the proposed RSCDNet model due to synthetic perturbations for Dataset - IV .

TABLE I
Comparison of Average Quality Metrics for Test Images on Dataset I With the Proposed RSCDNet Architecture and ITS Intermediate Stages

Metrics	Skip	ASPP	GL-ASPP	MSA	F1	Precision	Recall	Kappa	IoU	OA	TPR	TNR	FPR	FNR
U-Net [12] without preprocessing	✓	×	×	×	0.961	0.374	0.543	0.523	0.721	0.454	0.544	0.391	0.609	0.456
U-Net [12] with preprocessing	✓	×	×	×	0.972	0.743	0.851	0.842	0.773	0.948	0.852	0.955	0.045	0.148
Intermediate stage I	×	×	×	×	0.983	0.771	0.871	0.864	0.945	0.803	0.876	0.733	0.267	0.124
Intermediate stage II	✓	×	×	×	0.981	0.784	0.882	0.873	0.821	0.939	0.883	0.949	0.051	0.117
Intermediate stage III (4-8-12-16 rates)	×	✓	×	×	0.991	0.962	0.982	0.978	0.957	0.989	0.980	0.986	0.014	0.020
Intermediate stage III (6-12-18-24 rates)	×	✓	×	×	**0.993**	0.962	0.978	0.981	**0.973**	0.991	0.978	0.993	0.007	0.022
Proposed RSCDNet model (Final stage)	×	×	✓	✓	0.985	**0.982**	**0.984**	**0.983**	0.965	**0.995**	**0.983**	**0.994**	**0.006**	**0.017**

To understand the operations behind the enhanced performance of the MSA, a visual insight through heatmaps is presented. As can be seen in Fig. 6, the reduced intensity of the background after the MSA block signifies the increased confidence of the model in discerning the changed area from the unchanged area. Upon close observation of the changed areas after the GL-ASPP block, a further reduction f neighborhood intensities to yellow is a proof of the dilated convolutions with different rates increase the field of view of each pixel; this confirms its ability to accurately demarcate the boundaries of the changed areas from the background. It can be clearly concluded from the above presented discussion that the addition of MSA and GL-ASPP block resulted in a spike in the performance metrics.

C. Discussion

The proposed RSCDNet and its intermediate stages' quality metrics are compared to those of existing object segmentation networks, like U-Net [12] and other state-of-art models such as Siamese-based model [11], FDCNN [29], Triplet loss based model [30], DSMCSN [27], AGCDetNet [26] and UNet++ [25]. The developed RSCDNet model does not incorporate any pre-trained weights, rather it is trained in an end-to-end fashion. F1-score, precision, recall, kappa coefficient, and Jaccard score, True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), and False Negative Rate (FNR) are the metrics employed to assess the performance of the proposed

TABLE II
Comparison of Average Quality Metrics for Test Images on Dataset Ii With the Proposed Rscdnet Architecture and Its Intermediate Stages

Metrics	Skip	ASPP	GL-ASPP	MSA	F1	Precision	Recall	Kappa	IoU	OA	TPR	TNR	FPR	FNR
U-Net [12] without preprocessing	✓	×	×	×	0.652	0.531	0.844	0.643	0.478	0.967	0.839	0.967	0.033	0.161
U-Net [12] with preprocessing	✓	×	×	×	0.941	0.867	0.715	0.773	0.641	0.967	0.711	0.985	0.015	0.289
Intermediate stage I	×	×	×	×	0.981	0.973	0.967	0.978	0.967	0.991	0.967	0.989	0.011	0.033
Intermediate stage II	✓	×	×	×	0.981	0.979	0.978	0.978	0.972	0.989	0.978	0.991	0.009	0.022
Intermediate stage III (4-8-12-16 rates)	×	✓	×	×	0.978	0.967	0.978	0.971	0.961	0.989	0.978	0.989	0.011	0.022
Intermediate stage III (6-12-18-24 rates)	×	✓	×	×	0.978	0.979	0.980	0.981	**0.974**	0.990	0.982	0.990	0.010	0.018
Proposed RSCDNet model (Final stage)	×	×	✓	✓	**0.984**	**0.982**	**0.984**	**0.985**	0.961	**0.994**	**0.983**	**0.994**	**0.006**	**0.017**

TABLE III
Comparison of Average Quality Metrics for Test Images on Dataset Iii With the Proposed Rscdnet Architecture and Its Intermediate Stages

Metrics	Skip	ASPP	GL-ASPP	MSA	F1	Precision	Recall	Kappa	IoU	OA	TPR	TNR	FPR	FNR
U-Net [12] without preprocessing	✓	×	×	×	0.264	0.289	0.244	0.221	0.154	0.924	0.235	0.956	0.044	0.765
U-Net [12] with preprocessing	✓	×	×	×	0.781	0.867	0.711	0.767	0.643	0.971	0.706	0.989	0.011	0.294
Intermediate stage I	×	×	×	×	0.815	0.822	0.889	0.843	0.741	0.978	0.889	0.990	0.010	0.111
Intermediate stage II	✓	×	×	×	0.821	0.742	0.922	0.814	0.704	0.972	0.916	0.965	0.035	**0.084**
Intermediate stage III (4-8-12-16 rates)	×	✓	×	×	0.792	0.813	0.772	0.751	0.643	0.967	0.769	0.978	0.022	0.231
Intermediate stage III (6-12-18-24 rates)	×	✓	×	×	0.678	0.771	0.614	0.656	0.514	0.956	0.607	0.991	0.009	0.393
Proposed RSCDNet model (Final stage)	×	×	✓	✓	**0.884**	**0.872**	**0.895**	**0.873**	**0.781**	**0.993**	**0.894**	**0.995**	**0.005**	0.106

TABLE IV
Comparison of Average Quality Metrics for Test Images on Dataset Iv With the Proposed Rscdnet Architecture and Its Intermediate Stages

Metrics	Skip	ASPP	GL-ASPP	MSA	F1	Precision	Recall	Kappa	IoU	OA	TPR	TNR	FPR	FNR
U-Net [12] without preprocessing	✓	×	×	×	0.481	0.391	0.622	0.462	0.323	0.951	0.618	0.961	0.039	0.382
U-Net [12] with preprocessing	✓	×	×	×	0.691	0.741	0.652	0.681	0.531	0.982	0.652	0.991	0.009	0.348
Intermediate stage I	×	×	×	×	0.521	0.443	0.655	0.504	0.363	0.962	0.651	0.969	0.031	0.349
Intermediate stage II	✓	×	×	×	0.642	0.581	0.723	0.634	0.482	0.967	0.717	0.978	0.022	0.283
Intermediate stage III (4-8-12-16 rates)	×	✓	×	×	0.504	0.489	0.514	0.482	0.331	0.964	0.506	0.978	0.022	0.494
Intermediate stage III (6-12-18-24 rates)	×	✓	×	×	0.541	0.502	0.583	0.521	0.373	0.956	0.578	0.967	0.033	0.422
Proposed RSCDNet model (Final stage)	×	×	✓	✓	**0.752**	**0.741**	**0.762**	**0.744**	**0.603**	**0.984**	**0.765**	**0.991**	**0.009**	**0.235**

TABLE V
Comparison of Average Quality Metrics for Test Images on Dataset I With the Proposed Rscdnet Architecture and Benchmark Models

Metrics	F1 Score	Precision	Recall	Kappa	IoU	OA	TPR	TNR	FPR	FNR
Siamese (2017) [11]	0.856	0.878	0.851	0.661	0.634	0.970	0.852	0.978	0.022	0.148
UNet++ (2019) [25]	0.872	0.989	0.771	0.863	0.767	0.978	0.771	0.989	0.011	0.229
Triplet loss based model (2019) [30]	0.889	0.861	0.933	0.889	0.832	0.971	0.933	0.978	0.022	0.067
FDCNN (2020) [29]	0.542	0.501	0.589	0.525	0.373	0.952	0.589	0.967	0.033	0.411
DSMSCN (2020) [27]	0.434	0.401	0.482	0.351	0.478	0.856	0.478	0.911	0.089	0.522
AGCDetNet (2021) [26]	0.878	0.952	0.822	0.878	0.789	0.989	0.821	0.990	0.010	0.179
Proposed RSCDNet model	**0.982**	**0.983**	**0.981**	**0.984**	**0.963**	**0.992**	**0.985**	**0.991**	**0.009**	**0.015**

TABLE VI
Comparison of Average Quality Metrics for Test Images on Dataset Ii With the Proposed Rscdnet Architecture and Benchmark Models

Metrics	F1 Score	Precision	Recall	Kappa	IoU	OA	TPR	TNR	FPR	FNR
Siamese (2017) [11]	0.689	0.652	0.751	0.678	0.533	0.981	0.751	0.983	0.017	0.249
UNet++ (2019) [25]	0.763	0.842	0.701	0.762	0.622	0.978	0.706	0.985	0.015	0.294
Triplet loss based model (2019) [30]	0.967	0.955	0.978	0.973	0.941	0.975	0.975	0.981	0.205	0.205
FDCNN (2020) [29]	0.424	0.570	0.341	0.411	0.271	0.967	0.344	0.989	0.011	0.658
DSMSCN (2020) [27]	0.761	0.722	0.803	0.755	0.611	0.978	0.801	0.978	0.022	0.199
AGCDetNet (2021) [26]	0.963	0.934	0.978	0.945	0.920	0.989	0.978	0.989	0.011	0.022
Proposed RSCDNet model	**0.982**	**0.984**	**0.983**	**0.981**	**0.960**	**0.992**	**0.984**	**0.993**	**0.007**	**0.016**

RSCDNet model quantitatively. A qualitative comparison of the proposed RSCDNet model with other segmentation approaches and state-of-art architectures for the task of change detection is presented in Fig. 7.

Tables V, VI, VII and VIII contain tabulated results of the performance metrics of all the models. The evaluation was conducted by inferring the models at convergence. For the FDCNN model, in accordance in [29] Min Zhang et al., we used pre-trained VGG16 weights optimized on the aerial image data (AID), [42] and the exact implementation was followed. For Dataset II, III, and IV, our model provides a 2%, 3%, and 3% increment in F1-score respectively in contrast to the latest state-of-art model, while for Dataset I, a heavy increment in all the metrics can be seen. Fig. 7 reveals that the principal challenge in the generation of change maps is the variation in shapes and close boundaries of the changed areas. The proposed RSCDNet model has a better track record compared to other existing architectures in overcoming this challenge. The enhancement of the results due to the pre-processing unit can be inferred from the change maps produced with and without pre-processing in U-Net (base

TABLE VII
Comparison of Average Quality Metrics for Test Images on Dataset Iii With the Proposed Rscdnet Architecture and Benchmark Models

Metrics	F1 Score	Precision	Recall	Kappa	IoU	OA	TPR	TNR	FPR	FNR
Siamese (2017)[11]	0.631	0.684	0.532	0.567	0.462	0.905	0.531	0.956	0.044	0.469
UNet++ (2019) [25]	0.443	0.567	0.362	0.413	0.282	0.941	0.356	0.981	0.019	0.644
Triplet loss based model (2019) [30]	0.721	0.867	0.622	0.714	0.571	0.973	0.567	0.989	0.011	0.433
FDCNN (2020) [29]	0.582	0.522	0.651	0.562	0.413	0.961	0.652	0.967	0.033	0.348
DSMSCN (2020) [27]	0.854	0.855	0.852	0.846	0.761	0.945	0.851	0.971	0.029	0.149
AGCDetNet (2021) [26]	0.851	0.845	0.863	0.855	0.752	0.981	0.856	0.981	0.019	0.144
Proposed RSCDNet model	**0.881**	**0.874**	**0.890**	**0.871**	**0.783**	**0.989**	**0.891**	**0.992**	**0.008**	**0.109**

TABLE VIII
Comparison of Average Quality Metrics for Test Images on Dataset Iv With the Proposed Rscdnet Architecture and Benchmark Models

Metrics	F1 Score	Precision	Recall	Kappa	IoU	OA	TPR	TNR	FPR	FNR
Siamese (2017) [11]	0.441	0.333	0.651	0.412	0.278	0.943	0.654	0.951	0.049	0.356
UNet++ (2019) [25]	0.473	0.682	0.361	0.461	0.311	0.967	0.366	0.989	0.011	0.634
Triplet loss based model (2019) [30]	0.702	0.755	0.661	0.684	0.541	0.978	0.663	0.989	0.011	0.337
FDCNN (2020) [29]	0.378	0.304	0.513	0.352	0.234	0.952	0.508	0.956	0.044	0.492
DSMSCN (2020) [27]	0.567	0.414	0.903	0.545	0.396	0.943	0.895	0.943	0.057	0.105
AGCDetNet (2021) [26]	0.717	0.727	0.723	0.725	0.565	0.979	0.718	0.986	0.014	0.282
Proposed RSCDNet model	**0.754**	**0.743**	**0.765**	**0.743**	**0.602**	**0.983**	**0.764**	**0.992**	**0.008**	**0.236**

model). Siamese [11] and FDCNN [29] architecture are unable to differentiate between semantic and noisy changes due to low parameters thus resulting in higher false positives. Looking at the results of the Triplet loss based Model [30] and DSMSCN [27] model it can inferred that the models fail to demarcate change boudaries. On comparing the results of RSCDNet with the AGCDetNet model [26] and UNet++ [25], it can be concluded that processing the bi-temporal images separately for feature extraction rather than early image comparison helps to eliminate noise with the similar distribution. Currently, the AGCDetNet (2021) [26] is the best performing stat-of-art model. Several trials were conducted for the RSCDNet and AGCDetNet [26], the corresponding metrics like F1 score, Jaccard and Kappa are presented for comparison in Fig. 8. Median F-1 for RSCDNet is higher than that of AGCDetNet for all the datasets, thus implying that the RSCDNet is consistently stable. Thus, based on the extensive discussion and accompanying images, we can aver that the proposed model effectively resolves many of the issues that dogged previous architectures; the most pressing of which are concerns about change object shape and size, as well as border demarcation.

D. Robustness Experiments

In order to evaluate the robustness properties of the proposed RSCDNet model we performed a set of experiments with the synthetic perturbations. The bi-temporal images from the test set of Dataset IV were transformed with three different type of noise: Gaussian, Speckle and Poisson at three different intensities between 1 to 3; 1 being the lowest and 3 the highest. The noisy images are then passed through the already trained models on "clean" train images. The Fig. 9 plots the degradation in the F1-score for the state-of-art and the proposed RSCDNet model across all the noise variations. It can be observed that the proposed RSCDNet model shows minimal degradation in the performance. This can be attributed to the late feature comparison and shared encoders for the bi-temporal images. Looking closely at the performance and the images (Fig. 9 in supplementary paper) of the AGCDetNet [26], U-Net [12] and proposed RSCDNet model it can be see that early concatenation propagates the noisy changes through the model (U-Net) which are then suppressed by the spatial and channel attention blocks in AGCDetNet [26]. Going one step further in the RSCDNet model, building the trade-off between the dilated convolution and attention mechanism is helping the model to understand the noise distribution in the dataset.

TABLE IX
Comparison of Computational Complexity Analysis of Different Change Detection Models

	Parameters (in million)	FLOPs (in billion)	Avg. TT (in min)	PTPI (in ms)
Intermediate stage I	24.38	48.74	3.68	192
Intermediate stage II	25.77	51.5	4.01	176
Intermediate stage III	62.13	121.1	1.20	417
Proposed RSCDNet model	9.7	18.14	3.60	158
U-Net (2015) [12]	31.03	62.05	3.05	254
Siamese (2017) [11]	3.85	7.7	1.94	151
UNet++ (2019) [25]	10.19	20.38	5.48	550
Triplet loss model (2019) [30]	3.85	7.7	2.03	**140**
FDCNN (2020) [29]	**0.13**	**0.27**	**0.277**	1048
DSMSCN (2020) [27]	7.68	15.33	2.77	285
AGCDetNet (2021) [26]	44.1	88.2	3.07	650

E. Computational Complexity

The computational complexity of all models is examined with regards to the number of parameters used, floating point operations (FLOPs), training and prediction time per image as is presented in Table IX. Siamese model [11] utilizes the least parameters (3.35 million) and FLOPs (7.7 billion), while AGCDetNet [26] requires the most parameters (44.1 million) and 88.2 billion FLOPs amongst the state-of-art models. The proposed RSCDNet uses 9.7 million parameters and 18.14 billion FLOPs. It also employs shared weights in the dual encoder network, which aids in extracting similar features from bi-temporal images while also lowering the architecture's computational complexity. Although FDCNN model exhibits the least number of parameters and FLOPs but shows poor detection of the change areas. As is evident from Table IX, the proposed RSCDNet exhibits the least parameters, FLOPs and prediction time requirement among all other better performing state-of-the-art models, and hence, it is the most efficient solution to address the problems discussed in this paper.

VI. Conclusion

In remote sensing, change detection is a crucial step in natural resource investigation. An end-to-end deep learning architecture for change detection from high-resolution remote sensing imagery was presented in this paper. During the change detection task, the proposed RSCDNet resolved object shape and size heterogeneity as well as boundary-touching problems. As opposed to the base model U-Net, the proposed RSCDNet engaged a newly introduced, robust GL-ASPP block and modified self-attention mechanism to derive high-level spatial data and collectively hold vital information. Test outcomes on four public change detection datasets imply that our proposed RSCDNet model delivered excellent results in terms of F1, Precision, Recall, Jaccard, Kappa Coefficient, and other metric scores as compared to recent state-of-the-art change detection deep learning models. Further the fine trade-off between the complete context assimilation and field-of-view regulation is also providing the "robustness" to synthetic noise and minimal degradation in the performance. Currently the modified self-attention (MSA) module is taking a huge chunk of training time. Incorporating the current breakthroughs in self-attention for GPU training, the computational complexity of the design will be lowered and the architecture will be more efficient. Future work in the change detection domain will rely heavily on obtaining multi-spectral image samples and developing a dataset with distinct types of changes. The extension of this approach to instance segmentation of different kinds of change sets the stage for future work. Furthermore, with fine-tuning, the proposed model can be used for a multitude of other image segmentation domains.

Acknowledgment

The authors would thank all the editors and anonymous reviewers for their valuable suggestions and comments which helped to improve the quality of the research paper. Authors also would like to thank Ms.Sravya N and Prof.Fabio Dell'Acqua for reviewing and editing manuscript.

References

[1] M. K. Ridd and J. Liu, "A comparison of four algorithms for change detection in an urban environment," *Remote Sens. Environ.*, vol. 63, no. 2, pp. 95–100, 1998, doi: 10.1016/S0034-4257(97)00112-0.

[2] L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 3, pp. 1171–1182, May 2000, doi: 10.1109/36.843009.

[3] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and k-means clustering," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 772–776, Oct. 2009, doi: 10.1109/LGRS.2009.2025059.

[4] A. A. Nielsen, K. Conradsen, and J. J. Simpson, "Multivariate alteration detection (MAD) and MAF postprocessing in multispectral bitemporal image data: New approaches to change detection studies," *Remote Sens. Environ.*, vol. 64, no. 1, pp. 1–19, 1998, doi: 10.1016/S00344257(97)00162-4.

[5] S. Marchesi and L. Bruzzone, "ICA and kernel ICA for change detection in multispectral remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2009, vol. 2, pp. II-980–II-983, doi: 10.1109/IGARSS.2009.5418265.

[6] F. Bovolo, L. Bruzzone, and M. Marconcini, "A novel approach to unsupervised change detection based on a semisupervised SVM and a similarity measure," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 7, pp. 2070–2082, Jul. 2008, doi: 10.1109/TGRS.2008.916643.

[7] Z. Huang, X. Jia, and L. Ge, "Sampling approaches for one-pass land-use/land-cover change mapping," *Int. J. Remote Sens.*, vol. 31, pp. 1543–1554, 2010, doi: 10.1080/01431160903475399.

[8] D. Liu, K. Song, J. Townshend, and P. Gong, "Using local transition probability models in Markov random fields for forest change detection," *Remote Sens. Environ.*, vol. 112, pp. 2222–2231, 2008, doi: 10.1016/j.rse.2007.10.002.

[9] Z. Yetgin, "Unsupervised change detection of satellite images using local gradual descent," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 5, pp. 1919–1929, May 2012, doi: 10.1109/TGRS.2011.2168230.

[10] T. M., P. P., R. S. and A. T., "Land cover change detection using convolution neural network," in *Proc. IEEE 3rd Int. conf. Electron., Commun. Aerosp. Tech.*, 2019, pp. 791–794, doi: 10.1109/ICECA.2019.8821840.

[11] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep siamese convolutional network for optical aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1845–1849, Oct. 2017, doi: 10.1109/LGRS.2017.2738149.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.

[13] M. Volpi, D. Tuia, G. Camps-Valls, and M. Kanevski, "Unsupervised change detection by kernel clustering," in *Proc. SPIE*, 2010, pp. 264–271, doi: 10.1117/12.864921.

[14] M. Mignotte, "A fractal projection and Markovian segmentation-based approach for multimodal change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 8046–8058, Nov. 2020, doi: 10.1109/TGRS.2020.2986239.

[15] R. Touati, M. Mignotte, and M. Dahmane, "Multimodal change detection in remote sensing images using an unsupervised pixel pairwise-based Markov random field model," *IEEE Trans. Image Process.*, vol. 29, pp. 757–767, 2020, doi: 10.1109/TIP.2019.2933747.

[16] Z. Lv, T. Liu, and J. A. Benediktsson, "Object-oriented key point vector distance for binary land cover change detection using VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6524–6533, Sep. 2020, doi: 10.1109/TGRS.2020.2977248.

[17] S. Shi, Y. Zhong, J. Zhao, P. Lv, Y. Liu, and L. Zhang, "Land-use/land-cover change detection based on class-prior object-oriented conditional random field framework for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2000, Art. no. 5600116, doi: 10.1109/TGRS.2020.3034373.

[18] R. Touati, M. Mignotte, and M. Dahmane, "A circular invariant convolution model-based mapping for multimodal change detection," *Adv. Sci., Tech. Eng. Syst. J.*, vol. 5, pp. 1288–1298, 2020, doi: 10.25046/aj0505155.

[19] R. Touati, M. Mignotte, and M. Dahmane, "A reliable mixed-norm-based multi resolution change detector in heterogeneous remote sensing images," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3588–3601, Sep. 2019, doi: 10.1109/JSTARS.2019.2934602.

[20] S. Tian, Y. Zhong, A. Ma, and Z. Zheng, "Hi-UCD: A large-scale dataset for urban semantic change detection in remote sensing imagery," 2020, *arXiv:2011.03247*.

[21] S. Gopal and C. Woodcock, "Remote sensing of forest change using artificial neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 34, no. 2 pp. 398–404, Mar. 1996, doi: 10.1109/36.485117.

[22] L. Khelifi and M. Mignotte, "Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis," *IEEE Access*, vol. 8, pp. 126385–126400, 2020, doi: 10.1109/36.843009.

[23] R. Caye. Daudt, B. Le. Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proc. IEEE 25th Int. Conf. Image Process.*, 2018, pp. 4063–4067, doi: 10.1109/ICIP.2018.8451652.

[24] S. Sun, L. Mu, L. Wang, and P. Liu, "L-UNet: An LSTM network for remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2020, Art. no. 8004505, doi: 10.1109/LGRS.2020.3041530.

[25] D. Peng, Y. Zhang, and G. Wanbing, "End-to-end change detection for high resolution satellite images using improved UNet," *Remote Sens.*, vol. 11, 2019, Art. no. 1382, doi: 10.3390/rs11111382.

[26] K. Song and J. Jiang, "AGCDetNet: An attention-guided network for building change detection in high-resolution remote sensing images," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4816–4831, 2021, doi: 10.1109/JSTARS.2021.3077545.

[27] H. Chen, C. Wu, B. Du, and L. Zhang, "Deep siamese multi-scale convolutional network for change detection in multi-temporal VHR images," in *Proc. IEEE 10th Int. Workshop the Anal. Multitemp. Remote Sens. Images*, 2019, pp. 1–4, doi: 10.1109/Multi-Temp.2019.8866947.

[28] K. S. Basavaraju, N. Sravya, S. Lal, J. Nalini, C. S. Reddy, and F. Dell'Acqua, "UCDNet: A deep learning model for urban change detection from bi-temporal multispectral Sentinel-2 satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5408110, doi: 10.1109/TGRS.2022.3161337.

[29] M. Zhang and W. Shi, "A feature difference convolutional neural network-based change detection method," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7232–7246, Oct. 2020, doi: 10.1109/TGRS.2020.2981051.

[30] M. Zhang, G. Xu, K. Chen, M. Yan, and X. Sun, "Triplet-based semantic relation learning for aerial remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 266–270, Feb. 2019, doi: 10.1109/LGRS.2018.2869608.

[31] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[32] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[33] D. John and C. Zhang, "An attention-based U-net for detecting deforestation within satellite sensor imagery," *Int. J. Appl. Earth Observ. Geoinfo.*, vol. 107, 2022, Art. no. 102685, doi: 10.1016/j.jag.2022.102685.

[34] C. Benedek and T. Sziranyi, "Change detection in optical aerial images by a multilayer conditional mixed Markov model," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 10, pp. 3416–3430, Oct. 2009, doi: 10.1109/TGRS.2009.2022633.

[35] C. Benedek and T. Szirany, "A mixed Markov model for change detection in aerial photos with large time differences," in *Proc. IEEE 19th Int. Conf. Pattern Recognit.*, 2008, pp. 1–4, doi: 10.1109/ICPR.2008.4761658.

[36] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens.*, vol. 12, no. 10, 2020, Art. no. 1662, doi: 10.3390/rs12101662.

[37] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019, doi: 10.1109/TGRS.2018.2858817.

[38] K. Janocha and W. M. Czarnecki, "On loss functions for deep neural networks in classification," *Schedae Informaticae*, vol. 25, pp. 49–59, 2016, doi: 10.4467/20838476SI.16.004.6185.

[39] S. M. Pizer, R. E. Johnston, J. P. Ericksen, B. C. Yankaskas, and K. E. Muller, "Contrast-limited adaptive histogram equalization: Speed and effectiveness," in *Proc. 1st Conf. Visualiz. Biomed. Comput.*, 1990, pp. 337–345, doi: 10.1109/VBC.1990.109340.

[40] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," *Lect. Notes Comput. Sci.*, vol. 3408, pp. 345–359, 2005, doi: 10.1007/978-3-540-31865-1_25.

[41] N. Sravya, Priyanka, S. Lal, J. Nalini, C. S. Reddy, and F. Dell'Acqua, "DPPNet: An efficient and robust deep learning network for land cover segmentation from high-resolution satellite images," *IEEE Trans. Emerg. Top. Computat. Intell.*, early access, Jun. 24, 2022, doi: 10.1109/TETCI.2022.3182414.

[42] G. S. Xia et al., "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017, doi: 10.1109/TGRS.2017.2685945.