DRMNet: Difference Image Reconstruction Enhanced Multiresolution Network for Optical Change Detection

Avinash Chouhan, Arijit Sur, and Dibyajyoti Chutia, *Senior Member, IEEE*

***Abstract*—Change detection in satellite images is an important research area as it has a wide range of applications in natural resource monitoring, geo-hazard detections, urban planning, etc. Identifying physical changes on the ground and avoiding spurious changes due to other reasons like co-registration issues, change in illumination conditions, sun angle, and presence of cloud and fog is a challenging task. This work proposes a multitask learning based change detection model where two parallel pipeline architectures predict change map and image difference. The proposed model takes two images and their difference as input and provides them to a backbone network (BN). The output of the BN is fed into the proposed multiscale attention module for the effective identification of changes in multitemporal and very high-resolution aerial images. In another parallel path, the output of the BN is downsampled and passed to the proposed deconvolution with a subpixel convolution module to generate image difference. Two loss functions are utilized in two parallel paths to train the overall model in an end-to-end supervised setting. A comprehensive set of experiments have been carried out, and the results reveal that the proposed DRMNet model has achieved an F1 score improvement of 1.66% in CDD, 1.61% in SYSU, and 0.14% in LEVIR-CD datasets. It achieved an F1 score of 86.11% for the BCDD dataset with the new test image.**

***Index Terms*—Change detection (CD), difference image reconstruction, multiscale attention, optical remote sensing.**

I. Introduction

IDENTIFYING changes in bitemporal remotely sensed images is very useful in natural resource monitoring, urban planning, land monitoring, and other disaster mitigation applications. In the recent literature, plenty of change detection (CD) algorithms have been proposed. CD algorithms can be categorized into two groups based on data usage: homogeneous CD and heterogeneous CD. The homogeneous CD algorithm uses the input images taken from the same sensor, while the heterogeneous CD algorithm [1], [2] uses input images from different sensors to detect the changes. This work presents a homogeneous CD algorithm where the image pairs from the same sensor are used for analyzing the changes. Traditionally, arithmetic-based analysis [3], transformation-based analysis [4], and post-classification based algorithms [5], [6] are used for homogeneous CD, of which arithmetic-based analysis and transformation-based analysis are primarily unsupervised.

One of the main challenges of the CD task is determining the alteration at the pixel level. Recent studies proposed pixel-level CD methods [4], [7], [8] based on pixel differences between images taken in different time intervals for the same region. It maps the relationship between the same pixels between two different images to denote the changes in the spectral features. Object-level methods [9], [10] considered the group of pixels for the change identification. They try to identify the relation between pixels at the object level and consider that information for CD.

Deep learning (DL) based approaches offered the combination of pixel- and object-level approaches as they classify output at pixel level and considered semantic dependencies at the object level. For DL-based methods, two approaches are frequently used: early fusion networks and siamese networks. In an early fusion network [11]–[14], input images are first combined and then passed to the network to detect the changes, while in siamese networks [15]–[19], the input images are passed to the parallel stream of the network. The literature has observed that contextual information plays an important role in identifying the changes in high-resolution aerial images. To model the local context, large size kernels, local attention, and dilated convolutions are used [13], [17], [20]–[24]. Identification of long-range dependency between pixels is required to differentiate between actual and spurious changes, which are achieved using nonlocal operators and self-attention modules (SAMs) [15], [25]–[27].

In the literature, different approaches for CD are reported. Among them, adversarial learning based methods [28]–[30] for generating synthetic dataset, single image super-resolution based multiscale image generation method [31], and superpixels based siamese networks [32] become popular. Dense connectivity plays an important role in network design for CD in high-resolution aerial images. It provides feature reuse and alternative paths in the network. Recent network architectures for CD follow dense connections within the network [24], [33]. Preservation of the original resolution of features also becomes prominent to

Manuscript received January 25, 2022; revised March 18, 2022 and April 13, 2022; accepted April 26, 2022. Date of publication May 13, 2022; date of current version May 30, 2022. *(Corresponding author: Avinash Chouhan.)*

Avinash Chouhan is with the North Eastern Space Applications Centre, Meghalaya 793103, India, and also with the Department of Computer Science and Engineering, Indian Institute of Technology, Guwahati 781039, India (e-mail: avinash.chouhan@nesac.gov.in).

Arijit Sur is with the Department of Computer Science and Engineering, Indian Institute of Technology, Guwahati 781039, India (e-mail: arijit@iitg.ac.in).

Dibyajyoti Chutia is with the North Eastern Space Applications Centre, Meghalaya 793103, India (e-mail: d.chutia@nesac.gov.in).

Our model implementation is available at https://github.com/chouhan-avinash/DRMNet

Digital Object Identifier 10.1109/JSTARS.2022.3174780

avoid information losses and enhance the delineation process to identify actual changes.

The utilization of multispectral information is useful in the CD task [11], [34], [35]. For hyperspectral images, DL-based CD methods [36], [37] perform efficiently. CD can be modeled as a difference between input features. Attempts are made to replicate this in deep networks using the difference of feature maps [12], [33], [38] within the network. Postclassification-based methods [6] are also utilized for similar objectives but treat each image independently for classification.

It is observed in the literature that the temporal dependency of pixels plays a vital role in detecting the changes in images. For temporal dependence, a few recent schemes [35], [39], [40] have used recurrent neural networks. In addition, coregistration of image pairs and input normalization is also important for CD tasks as a minor shift in input data, variation of surrounding conditions, cloud, fog, shadow, and sun angle may create the pseudochange effects. These effects are required to nullify for efficient detection strategies.

In the satellite images, physical changes are categorized as spatial and temporal changes. A long-range context correlation between pixels is necessary for modeling such spatial and temporal changes. Capturing long-range context correlation is difficult due to convolutional kernels' limited field of view. Moreover, CD is a dense labeling problem, and, thus, the prediction of a precise change map for high-resolution satellite images is difficult. The other best published method [13] resolves these challenges using dilated convolution and self-attention. Preservation of the image's original resolution is also vital to avoid information loss due to downsampling and subsequent upsampling processes. Another best published method [24] handles this through the use of NestedUnet architecture that preserves original resolution feature maps. This work proposes an end-to-end multitasking architecture that predicts the change map and the image difference through two parallel architectures. One path predicts the change map by analyzing the multiresolution backbone with the help of a multiscale attention module (MSAM). The multiresolution backbone ensures the preservation of original resolution, and multiscale self-attention extracts long-range pixel dependencies required to capture finer details. The other path uses a deconvolution network using subpixel convolution to predict the accurate image difference from downsampled features generated from the backbone network (BN).

The rest of this article is organized as follows. Section II contains a comprehensive literature review for related work. The detailed architecture of the proposed model is described in Section III. The experimental setup is given in Section IV. Experimental results having a comparison of the proposed scheme with the existing state-of-the-art (SOTA) results are given in Section V. An ablation study is included in Section VI to highlight the usefulness of individual modules of the proposed model. Finally, Section VII concludes this article.

II. Literature Survey

CD requires a pair of images as input that a siamese-based convolutional neural network can effectively utilize. Zhan *et al.* [18] proposed a siamese neural network having increasing kernel size with weighted contrastive loss to handle data imbalance. Zhang *et al.* [19] pointed out that contrastive loss ignored semantic dependencies between pixels. To resolve it, the authors proposed the use of a siamese network that utilized enhanced triplet loss based on triplet selection of anchor features, positive features, and negative features. Peng *et al.* [14] followed a different approach than the siamese network and used concatenated input pairs. They proposed the use of UNet++ [41] as it utilized dense connections and multiscale feature streams with multiple side output fusion.

Attention-based models have also been proposed frequently for CD. Zhang *et al.* [20] utilized a double UNet based encoder–decoder network to produce coarse and refine scale predictions. Attention gates are used in coarse networks, and their output is enhanced in the second network using ground truth and residual connections. In another attention-based model, Zhang *et al.* [21] proposed an image fusion network (IFN) that used two fully convolutional networks for extraction of features that are passed to a difference discrimination network for change map generation. In a similar direction, Chen *et al.* [25] proposed the use of spatial and temporal attention in a deep network to identify bitemporal changes. Besides these, channelwise attention modules and atrous convolutions are used by Song *et al.* [13] to generate multiscale and multicontext features. In a similar line of thought, bilateral semantic fusion siamese network (BSFNet) is proposed by Du *et al.* [17] which uses the channel and spatial attention. Other significant attention-based method includes ADS-Net by Wang *et al.* [22] and DTCDSCN by Liu *et al.* [23].

Time-varying sequential inputs are also essential in CD to model the temporal dependency. Mou *et al.* [35] proposed a recurrent convolution neural network to utilize temporal features in addition to spatial and spectral features. The recurrent module analyzes the features extracted by CNN layers to model the temporal dependency. Chen *et al.* [39] introduced a siamese convolutional multiple-layer recurrent neural network that used extracted features from the convolutional neural network and recurrent neural network to find the changes. Ru *et al.* [42] proposed a correlation-based fusion module that utilized deep features to find the correlation between instances and fused it to find changes. This correlation is utilized for cross-temporal fusion.

The difference of input pair depicts an abstract representation of the change that happened between inputs. Peng *et al.* [12] introduced a dense attention module that used multiple upsampling attention units for feature fusion. They also proposed the use difference of input pairs in the enhancement unit. Zang *et al.* [38] proposed a feature difference convolution neural network that used shared pretrained weight for feature extraction, and the decoder module utilized the difference of extracted features. Zhang *et al.* [33] proposed DifUnet++ that utilized a difference pyramid at multiple scales of features with Unet++[41] as a base model and used learning-based Dupsampling instead of bilinear upsampling.

Generative adversarial networks are a promising area of research for an image-to-image translation task. Zhao *et al.* [28] proposed attention gates generative adversarial adaptation network that used attention gates for spatial constraint with domain similarity loss for multiple CD. Chen *et al.* [29] proposed instance-level change augmentation (IAug) to produce a

synthetic building change instance dataset. Change detection network (CDNet) utilized actual and augmented building instances for CD. Liu *et al.* [31] introduced a super-resolution based change detection network (SRCDNet) which used adversarial learning for the generation of high-resolution images. These multiscale images are passed to feature extractors with attention modules to produce multiscale CD maps. Zhang *et al.* [32] proposed a superpixel enhance CD network, which is based on two parallel siamese networks of superpixel subsampling networks. A deep network utilized extracted superpixels, and an adaptive superpixels merging module is proposed. Lebedev *et al.* [30] utilized Pix2Pix, a generative adversarial network, to produce a change map and also introduced CDD dataset.

Fang *et al.* [24] proposed densely connected siamese network SNUNet-CD, which used NestedUNet based siamese network. Ensemble channel attention module is proposed which aggregates and refines the features from multiple levels of the network. Foivis *et al.* [26] proposed fractal Tanimoto similarity metric and FracTAL ResNet block. Fractal Tanimoto attention layer is proposed for improvement in scale-dot based self-attention mechanism. The transformer is an encoder–decoder based network that utilizes multihead self-attention. Chen *et al.* [27] proposed a bitemporal image transformer (BiT) that used token-based context encoding. A transformer decoder is used for the conversion of taken-based encoding into pixel-based output. Shi *et al.* [43] proposed a deep supervised attention metric based network (DSAMNet), which used siamese network and channel attention. These schemes are very efficient and show SOTA performance for different datasets. Shi *et al.* [43] mention four standard datasets, namely, CDD dataset by Lebedev *et al.* [30], LEVIR-CD dataset by Chen *et al.* [25], BCDD dataset by Ji *et al.* [44], and SYSU dataset by Shi *et al.* [43]. Most of the current schemes have used these datasets for experimentation.

From the reviewed literature, we concluded that following are the major challenges faced by existing architectures.

1) Encoder–decoder based network suffers information loss due to downsampling of the original resolution.
2) Class imbalance problem may occur due to the insufficient number of change classes.
3) Distinction between actual and spurious changes is difficult.

III. Proposed Scheme

A. Motivation

In this work, we have proposed a dense high-resolution network to mitigate the shortcomings mentioned above in the existing literature. Two input images (of which the changes are detected) and the modulus of their difference are concatenated and treated as input in this work. This input is fed to the BN to preserve the original resolution and extract high-resolution features. Apart from the BN modules, the proposed model has two other significant modules, named MSAM and deconvolution with subpixel convolution module (DSCM). MSAM models the long-range dependencies between pixels by using self-attention maps. DSCM helps to remove the spurious changes effectively to produce a more accurate image difference using a deconvolution layer through subpixel convolution. In addition, two different loss functions are used to tackle the class imbalance problem. The main contributions of this work are as follows.

1) The proposed DRMNet is a multitasking bitemporal CD model that can efficiently model the long-range pixel dependencies for a very high-resolution aerial image using an MSAM.
2) The DRMNet also incorporates a DSCM to get more precise change maps by enhancing the quality of feature representation generated by the BN.
3) In the DRMNet, a modulus difference-based loss function is used to detect the changes more precisely by the DSCM.

To justify the efficacy of the proposed scheme over the existing SOTA literature, we use CDD, SYSU, LEVIR-CD, and BCDD datasets for benchmark evaluation. Experimental results reveal that the proposed scheme outperforms the current SOTA methods. An ablation study is also included at the end to justify the contributions of the various architecture modules and the impacts of different hyperparameters on the proposed model.

Algorithm 1: An algorithm for DRMNet training. Here I_1, I_2 are input images of two different timestamps, $|I_1 - I_2|$ is the modulus of input pair difference, GT is the ground truth, O_1 is the final change map, O_2 is auxiliary output, and L_1, L_2 are calculated loss values.

Data: $Sample_{train}$ = (I_1, I_2, GT) $\in$ Trainset
Result: O_1
while *exist(*$Sample_{train}$*)* **do**
  $I \leftarrow concat(I_1, I_2, |I_1 - I_2|)$;
  $F \leftarrow BN(I)$;
  $O_1 \leftarrow MSAM(F)$;
  $F_{downsample} \leftarrow Downsampling(F)$;
  $O_2 \leftarrow DSCM(F_{downsample})$;
  $L_1 \leftarrow Loss_1(O_1, GT)$;
  $L_2 \leftarrow Loss_2(O_2, |I_1 - I_2|)$;
end

B. Proposed Network Architecture

The proposed DRMNet model has three primary modules, as depicted in Fig. 1. The first one is BN, which is an extension of an existing model named HRNet [45]. The second is the MSAM to generate self-attention at different scales. The third one is DSCM, which is used to predict the difference (of two input images) reconstruction. The proposed DRMNet model takes two images (i.e., image *A* and image *B*, where we have to detect changes between *A* and *B*) and the corresponding image difference ($|A - B|$) as input. First, the concatenation of three images (*A*, *B*, and $|A - B|$) is passed through a sequence of convolution, batch normalization, and ReLU layers, and then its output is fed into BN. This module outputs 48 channel feature matrices of dimension $N \times N \times 48$ where $N \times N$ is the input image(s) dimension. We selected 48 channels' output after initial layers based on the ablation study of different initial channel configurations as presented in Table VI. The output

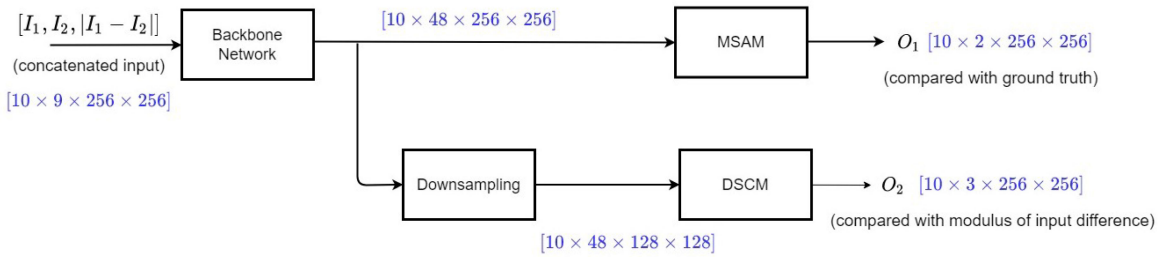

Fig. 1. DRMNet complete architecture. MSAM is a multiscale attention module, and DSCM is a deconvolution with a subpixel convolution module. Here, I_1 and I_2 are inputs passed to network and produced outputs represented as O_1 and O_2. The feature's shape at each stage of the network is represented with blue color.

TABLE I
CDD DATASET PERFORMANCE COMPARISON

	Precision	Recall	F1 Score	IoU	OA
ADS-Net [22]	89.79	79.58	82.72		
UNet++ [14]	89.54	87.11	87.56		
IFN [21]	94.96	86.08	90.30		97.71
BA^{2}Net [20]	88.12	95.28	91.36		98.94
BSFNet [17]	90.5	93.3	91.9		98.10
DASNet [15]	93.2	92.2	92.7		98.2
SRCDNet [31]	-	-	92.94		
DiffUNet++ [33]	92.15	94.63	93.37		
DSAMNet [43]	94.54	92.77	93.69	88.13	
DDCNN [12]	96.71	92.32	94.46	89.51	98.64
SNUNet [24]	96.3	96.2	96.2		
LSS-Net [48]	96.74	95.87	96.30	-	-
AGCDetNet [13]	95.03	98.10	96.54	-	99.13
Ours	97.92	98.49	98.20	96.46	99.57

TABLE II
LEVIR-CD DATASET PERFORMANCE COMPARISON

	Precision	Recall	F1 Score	IoU	OA
BSFNet [17]	82.7	94.0	88.0	-	97.0
STANet [25]	83.8	91.0	87.3		
CDNet + IAug [29]	91.6	86.5	89.0	-	-
DiffUNet++ [33]	92.4	87.1	89.6	-	-
ADS-Net [22]	89.67	91.36	89.80	-	-
BiT [27]	89.24	89.37	89.31	80.68	98.92
SNUNet [24]	90.61	89.01	89.80	81.49	98.97
DDCNN [12]	91.85	88.69	90.24	82.21	98.11
AGCDetNet [13]	92.12	89.45	90.76	83.09	-
CEECNet [26]	93.81	89.92	91.83	84.89	-
Ours	93.05	90.91	91.97	85.13	99.19

TABLE III
SYSU DATASET PERFORMANCE COMPARISON

	Precision	Recall	F1 Score	IoU	OA
FC-EF [11]	74.32	75.84	75.07	60.09	
BiDateNet [16]	81.84	72.60	76.94	62.52	
STANet [25]	70.76	85.33	77.37	63.09	
DSAMNet [43]	74.81	81.86	78.18	64.18	
SNUNet [24]	78.16	79.68	78.92	65.18	89.96
Ours	84.55	76.86	80.53	67.39	91.23

TABLE IV
WHU BCDD DATASET PERFORMANCE COMPARISON

	Precision	Recall	F1 Score	IoU	OA
*BiT [27]	86.64	81.48	83.98	72.39	98.75
*DDCNN [12]	91.85	88.69	90.24	82.21	98.11
*AGCDetNet [13]	92.12	89.45	90.76	83.09	-
*CEECNet [26]	95.57	92.043	93.77	88.23	-
*FCCDN [49]	96.39	91.24	93.73	88.20	-
*DTCDSCN [23]	-	89.32	89.01	78.08	-
*LSS-Net [48]	94.18	93.36	93.77	-	-
SNUNet [24]	85.25	81.09	83.13	71.12	98.80
Ours	87.93	84.37	86.11	75.61	99.01

Here, * represents results computed on different split for test data and not comparable

TABLE V
ABLATION STUDY FOR LOSS PARAMETER VALUE

α	0.5	0.6	0.7	0.8	**0.9**	1
F1 Score	91.09	91.15	91.18	91.29	**91.46**	91.43
IoU	83.56	83.77	83.89	84.19	**84.28**	84.25

TABLE VI
ABLATION STUDY FOR NUMBER OF INITIAL CHANNELS TO BE USED

	Parameters	Precision	Recall	F1 Score
Base-32	15.67 M	90.01	86.99	89.11
Base-48	34.94 M	90.67	88.79	89.72
Base-64	61.70 M	90.70	88.78	89.73

Parameters are calculated in millions (M)

is fed to the MSAM module to identify whether the image is changed or not, and the inference is tallied with ground truth to find the corresponding loss. The exact output from the BN is downsampled, and the downsampled version is fed into the DSCM. The output of the DSCM is compared with the input images differences to generate the corresponding loss (L2).

C. Backbone Network Module

This module extracts the high-resolution image features while preserving the original image resolution. The block diagram of the BN architecture is depicted in Fig. 2. We used residual block as the basic unit for this backbone. Residual connections can be represented as

$$R(x) = \kappa(x) + x. \tag{1}$$

Here, x is input to block with residual connection, κ is nonlinear mapping, and R is residual block output. We extended an existing network called HRNet proposed by Wang *et al.* [45]

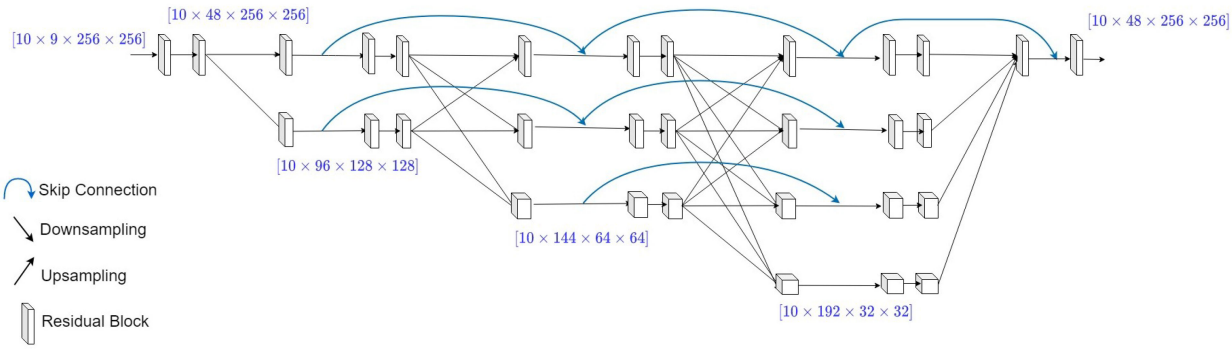

Fig. 2. Backbone module. Here, blue connections represent proposed skip connections between the same resolution stream. The down arrow represents downsampling, and the up arrow represents upsampling of features.

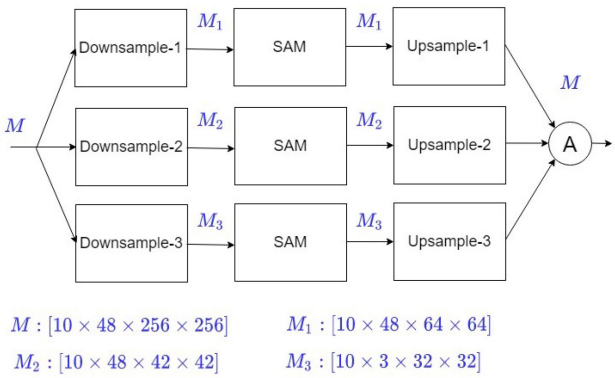

Fig. 3. Multiscale attention module. Here, SAM is the self-attention module. Downsample-1, Downsample-2, Downsample-3, and Upsample-1, Upsample-2, Upsample-3 represent downsampling and upsampling operations at different scales.

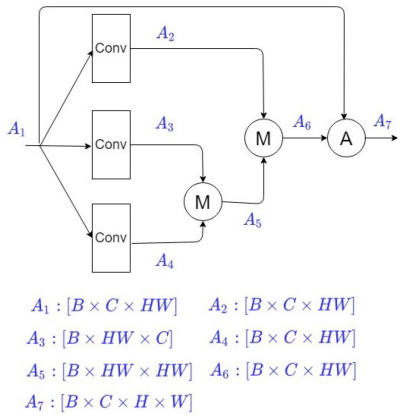

Fig. 4. Self-attention module, M represents multiplication operation, and A is addition operations.

where the high-resolution representations of the input image are maintained by connecting the high-to-low resolution convolution streams in parallel and by exchanging the information across resolutions frequently. Likewise, in this work, we generated multiple streams of different resolutions (e.g., 1, $\frac{1}{2}$, $\frac{1}{4}$, etc.) from the input image. These multiresolution streams are fully connected, as shown in Fig. 2. Features of different resolutions are combined using the addition operator. We extended the HRNet model by adding the skip connections between the same resolution stream for parameter sharing and better gradient flow. This BN module provides high-resolution feature extraction and multiresolution information fusion, which are required to identify actual changes in high-resolution images.

D. Multiscale Attention Module

As described in Fig. 1, the output of the BN is fed into two different modules in parallel. One is the MSAM, and the other is the DSCM. In this subsection, MSAM is described with an illustrative example as given in Fig. 3. The self-attention [46] is used at multiple lower scales to produce aggregated multiscale self-attentive features. This module essentially generates self-attention maps at different lower scales such as $\left[\frac{1}{4}, \frac{1}{6}, \frac{1}{8}\right]$ of the original resolution and fuses them as described in Fig. 3. The input (original resolution) features are downsampled in different lower resolutions ($\frac{1}{4}, \frac{1}{6}, \frac{1}{8}$) and passes through the SAM individually. Then all three SAM responses are upsampled with respect to their original resolutions and then added to get the final response. The self-attention is used to find the long-range dependencies between pixels. The block diagram of SAM has depicted in Fig. 4. Self-attention captures the global interaction of the pixels using dot product operation between the linear representation of inputs. For the given input I, it can be represented as

$$Z = f(\theta(I), \psi(I)) * \mu(I) + I \tag{2}$$

where Z is self-attentive output map, f is mapping function used in self-attention, and $\theta(.), \psi(.), \mu(.)$ are linear functions. The linear functions are implemented using 1×1 convolutions. These linear functions produced a linear transformation of input I which is utilized by mapping function f and distributed over linear representation generated by $\mu(.)$. The function f uses multiplication and softmax operation and is defined as

$$f(\theta(I), \psi(I)) = \text{softmax}(\theta(I) * \psi(I)). \tag{3}$$

Creating a self-attention map in the original resolution is very memory extensive and not feasible for large image patches. In this module, we propose to generate a self-attention map at multiple lower resolutions and later combine the upsampled maps.

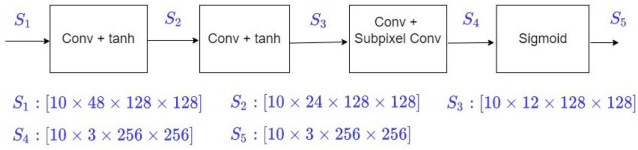

Fig. 5. Deconvolution with sub-pixel convolution module (DSCM). Here, Conv is the convolutional layer.

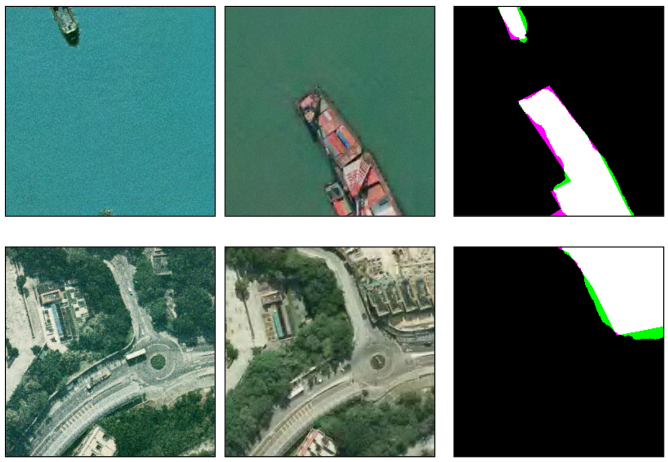

Fig. 6. Output visualization of 256×256 pixels patches for SYSU test dataset. Here, the white color is the actual change detected by the model, green color is pseudochange detected by the model, and the pink color is the actual change missed.

This approach produced better results than self-attention maps at single resolution and required less computational resources than self-attention at original resolution. The output of MSAM (Z_{final}) can be represented by the following equations:

$$Z_{\text{final}} = Up_{4x}\left(Z_{\frac{1}{4}}\right) + Up_{6x}\left(Z_{\frac{1}{6}}\right) + Up_{8x}\left(Z_{\frac{1}{8}}\right) \quad (4)$$

where $Z_{\frac{1}{4}}, Z_{\frac{1}{6}}$, and $Z_{\frac{1}{8}}$ are self-attentive maps, generated at $\frac{1}{4}$th, $\frac{1}{6}$th, and $\frac{1}{8}$th resolutions of input I. Up_{4x}, Up_{6x}, and Up_{8x} represent upsampling operation with upsample rates of 4, 6, and 8. It is used to produce the aggregated attentive map (Z_{final}) at original resolution.

E. Deconvolution With Subpixel Convolution Module

This module is a deconvolution (decoder) module, which takes a subsampled version of image difference features (generated by the BN module) and reconstructs the image difference in the original resolution. The resolution scaling (upsampling) is made through the subpixel convolution process. It has been experimentally observed that by predicting image difference from downsampled feature maps, this module can remove the spurious changes effectively to produce a more accurate image difference. Thus, the DSCM is one of the novel contributions of the proposed work. A basic block diagram of this DSCM is depicted in Fig. 5. The DSCM consists of two convolutions (with Tanh activation function) layers, one convolution layer with subpixel convolution, and a sigmoid layer. The subpixel convolution [47] is defined as a standard convolution in low-resolution space followed by a periodic shuffling operation. It is used as a part of the proposed deconvolution process for the required upsample process to reconstruct the image difference in the original resolution. Architecturally, this pipeline (BN module, downsampler and followed by DSCM as deconvolution process) forms an encoding–decoding like architecture which helps to remove the spurious changes (due to image acquisition, spatial de-synchronization, etc.) from the image difference (i.e., the change map). Subpixel convolution is a well-known technique which transforms the input of size $H \times W \times C$ to $(H \times d) \times (W \times d) \times \frac{C}{d^2}$ with H, W, and C being height, width, and channels of input and d is the upsampling factor. We have used the sigmoid activation function because it gives output in the range of [0, 1]. This module can be mathematically formulated as

$$O_2 = \sigma\Bigg(PS\Bigg(w_3\Bigg(\tanh \times \Bigg(w_2\Bigg(\tanh\left(w_1\left(O_F\frac{1}{2}\right) + b_1\right) + b_2\Bigg) + b_3\Bigg)\Bigg)\Bigg)\Bigg). \quad (5)$$

Here, PS is pixel shuffle convolution, $O_F\frac{1}{2}$ is downsampled output of backbone, w_i is weight of convolution kernel, and b_i is bias of convolution.

The output of this DSCM is compared to the modulus of input pair difference using the loss function (L_2), which is calculated as mean square error (mse). This loss function trains the BN module for better feature generation. The features generated from the BN module are downsampled and fed to DSCM to predict the change map by removing the spurious contents from the input features. To get the loss, these predicted change maps (O_2) are compared with the image differences (using the L_2 loss function).

F. Loss Function

In Fig. 1, it can be observed that two loss functions are used in the proposed model. First, loss function (L_1) is calculated between O_1 and ground truth (G). In this case, we have used a combined loss function as it has been used in SNUNet [24]. The authors have argued that there exists a sample imbalance effect as the number of unchanged pixels is often far more than the number of changed pixels. To reduce the sample imbalance effect, a combination of weighted entropy loss L_{ce} and dice loss L_d has been used. L_{ce} and L_d can be represented as per the following equations:

$$L_1 = L_{ce} + L_d \quad (6)$$

$$L_{ce} = -\sum_{1}^{N} c \times \log(S(m)_c) \quad (7)$$

$$L_d = 1 - \frac{2 \times O \times S(m)}{O + S(m)}. \quad (8)$$

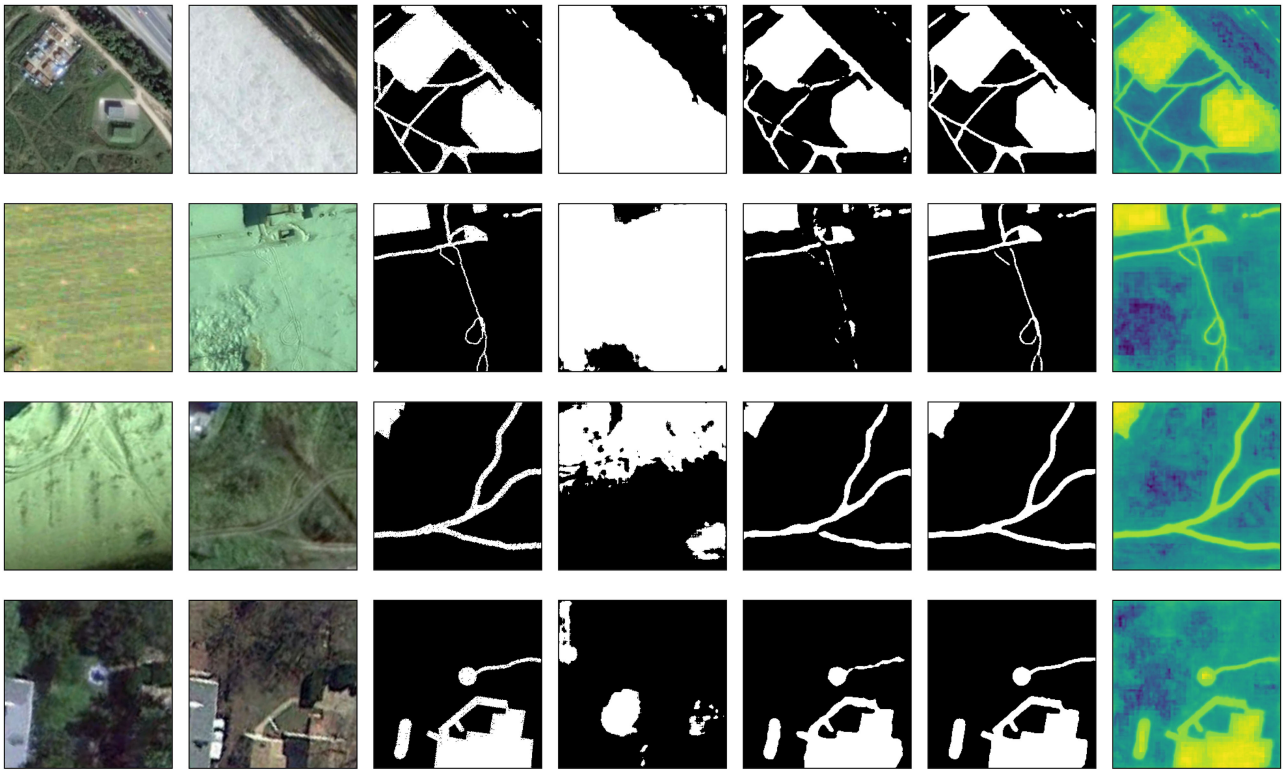

Fig. 7. 256×256 pixel patches output visualization for CDD test dataset. (a) Image 1. (b) Image 2. (c) Ground truth. (d) FC-EF output. (e) SNUNet output. (f) Our output. (g) Heat map.

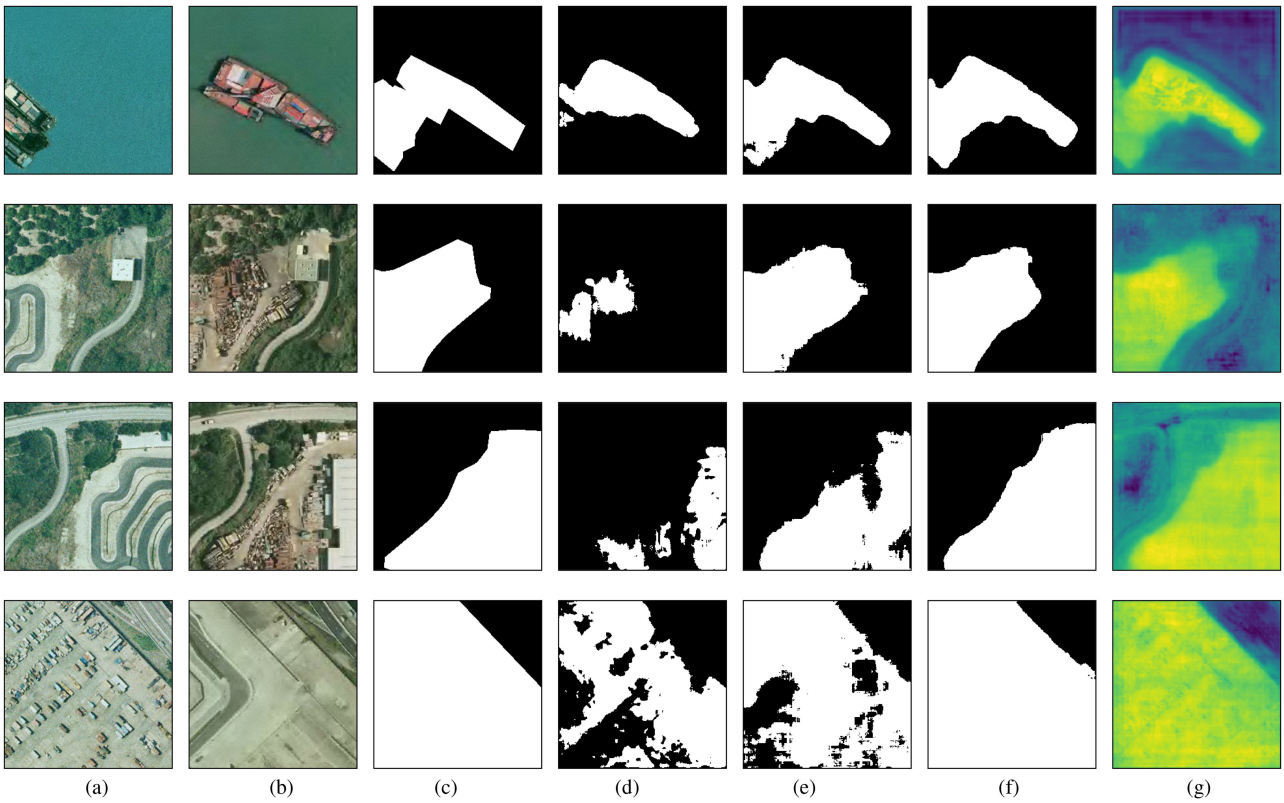

Fig. 8. 256×256 pixel patches output visualization for SYSU test dataset. (a) Image 1. (b) Image 2. (c) Ground truth. (d) FC-EF output. (e) SNUNet output. (f) Our output. (g) Heat map.

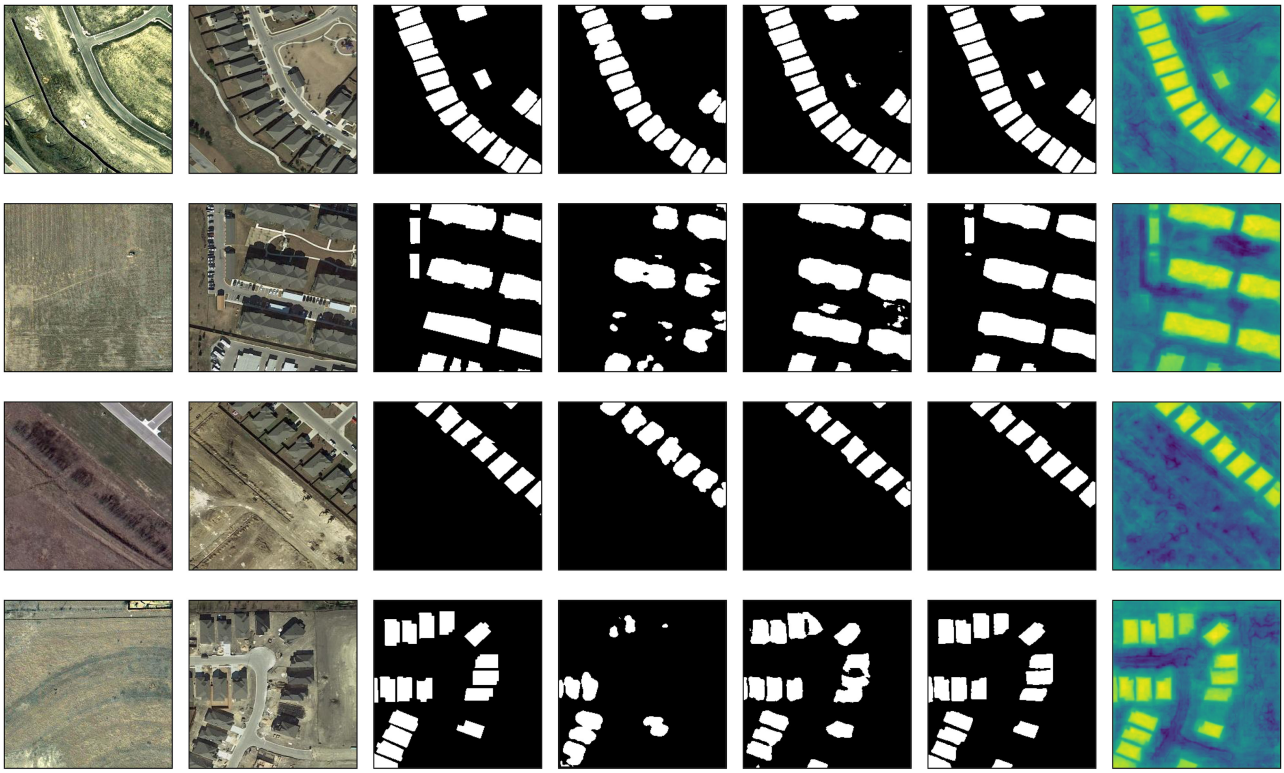

Fig. 9. 256×256 pixel patches output visualization for LEVIR test dataset. (a) Image 2. (b) Image 2. (c) Ground truth. (d) FC-EF output. (e) SNUNet output. (f) Our output. (g) Heat map.

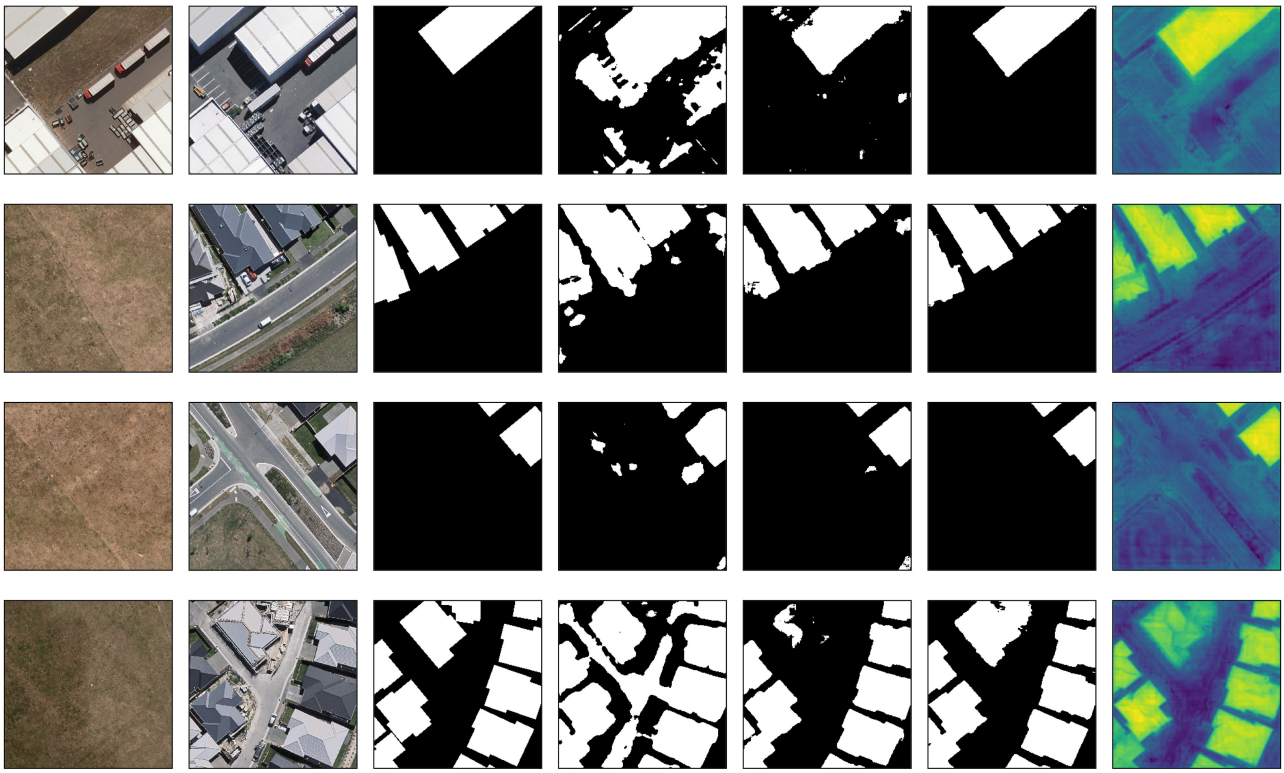

Fig. 10. 256×256 pixel patches output visualization for WHU test dataset. (a) Image 1. (b) Image 2. (c) Ground truth. (d) FC-EF output. (e) SNUNet output. (f) Our output. (g) Heat map.

Here, $S(.)$ is the softmax function, c is a class vector, m is the model's output before the softmax layer, and O is the ground truth.

The second loss function (L_2) is calculated as mse between O_2 and $|I_1 - I_2|$. The overall loss function is calculated as follows:

$$\text{Loss}(I_1, I_2, G, O_1, O_2) = L_1(G, O_1) + \alpha \times L_2(|I_1 - I_2|, O_2). \quad (9)$$

Here, L_{ce}, L_d, α, and L_{mse} are weighted cross-entropy loss, dice loss, loss weight constant (variable with value between [0, 1]), and mse loss. L_1 is calculated with the predicted change map using the output of MSAM (O_1) with the actual ground truth, and L_2 loss is calculated between the predicted image difference using the output of DSCM (O_2) and the ground truth. We find the optimal value of α using exhaustive experiments as listed in Table V. From this, we found that the α value 0.9 produces the optimal performance.

IV. Datasets

We have used the Google dataset CDD proposed by Lebdev *et al.* [30], LEVIR-CD dataset introduced by Chen *et al.* [25], WHU BCDD [44] dataset, and SYSU [43] dataset. CDD dataset consists of 16 000 images of 256×256 pixels containing RGB data with training, validation, and test set sizes of 10 000, 3000, and 3000, respectively. LEVIR-CD dataset contains 637 bitemporal images of size 1024×1024 pixels with RGB bands. We have generated 256×256 size patches from the LEVIR-CD dataset for our experiments. WHU BCDD dataset consists of a training image of size 15354×21243 pixels and a testing image with size 15354×11265 pixels. Earlier, a single image of size 15354×32507 was provided, and the test image was generated randomly. We cropped 256×256 size patches from the training image to train the model. SYSU dataset consists of 20 000 aerial images for CD with the train set, validation set, and test set divided into 12 000, 4000, and 4000 images.

CDD dataset contains changes of buildings, roads, etc. SYSU dataset contains changes of building, road, sea construction, vegetation, construction, etc. LEVIR-CD and WHU BCDD datasets focus on building CD.

V. Experiments

A. Training and Hyperparameters

We have used an initial learning rate of 0.001 and a batch size of 10 for the training of our network. We used the Nvidia P100 graphics card with 32 GB of graphics memory. We selected batch size based on this available graphics memory. We used data augmentations of horizontal and vertical flipping with 50% probability and random rotation to increase their variability. For each dataset, the model is trained for 300 epochs with early stopping when it cannot optimize further. Input pairs and ground truths are normalized before passing to the network. During inference, we used test time augmentation of horizontal and vertical flipping and 90° rotations. Final maps are produced using the average of all augmented outputs.

B. Comparative Analysis

We used precision (P_r), recall (R_c), F1 score (F_1), intersection over union (IoU), and overall accuracy (OA) as performance metrics for quantitative comparison of outputs. These metrics are used in recently published works for comparative study. These are calculated using true positive (t_p), false positive (f_p), true negative (t_n), and false negative (f_n) values as per following equations:

$$P_r = \frac{t_p}{t_p + f_p} \quad (10)$$

$$R_c = \frac{t_p}{t_p + f_n} \quad (11)$$

$$F_1 = 2 \times \frac{P_r \times R_c}{P_r + R_c} \quad (12)$$

$$\text{OA} = \frac{t_p + t_n}{t_p + f_p + t_n + f_n} \quad (13)$$

$$\text{IoU} = \frac{t_p}{t_p + f_p + f_n}. \quad (14)$$

Numerical comparisons against the SOTA methods on different datasets are presented in Tables I–IV with the best result highlighted in the red color. As it can be observed in Table I, for the CDD dataset, our model achieved 99.57% OA with F1 score of 98.20 which are 0.44% and 1.66%, respectively, higher than the SOTA result [13]. The corresponding subjective (visual) results are presented in Fig. 7. As per Table II and Fig. 12, it is observed that our model achieved *F1 score* of 91.97% which is 0.14% higher than the SOTA result [26] for LEVIR-CD dataset. From Table III and Fig. 8, it is observed that against the SYSU dataset, we achieved 91.23% OA and 80.53% *F1 score* which are 1.27% and 1.61% higher, respectively, than the SOTA result [43]. For the BCDD dataset, we reported our result on a separate test image provided in dataset. Earlier approaches have reported results on the random test split. These results are not comparable with our output. For comparison against the BCDD dataset, we have trained SNUNet [24] and generated results on the test image. Our results are presented in Table IV where not comparable results are shown with * mark and visual comparison shown in Fig. 10.

In Table V, detailed experiment results are presented for finding the optimal value of α. We trained our final model on the LEVIR-CD dataset. The experiment started with an initial value of 0.5 for α and achieved F1 score of 91.09% and IoU of 83.56%. We repeated this experiment with an increase of 0.1 in the α value till the final value of 1.0. After analyzing the result produced, we found that 0.9 is the optimal value.

C. Visual Analysis

Visual test results are shown in Fig. 6 using RGB composite (ground truth in red and blue bands and prediction in green band) images to demonstrate the accurate prediction capability of the proposed model. We can conclude from these figures that our model produced results very similar to actual ground truth. It missed a small number of actual changes with few detection of

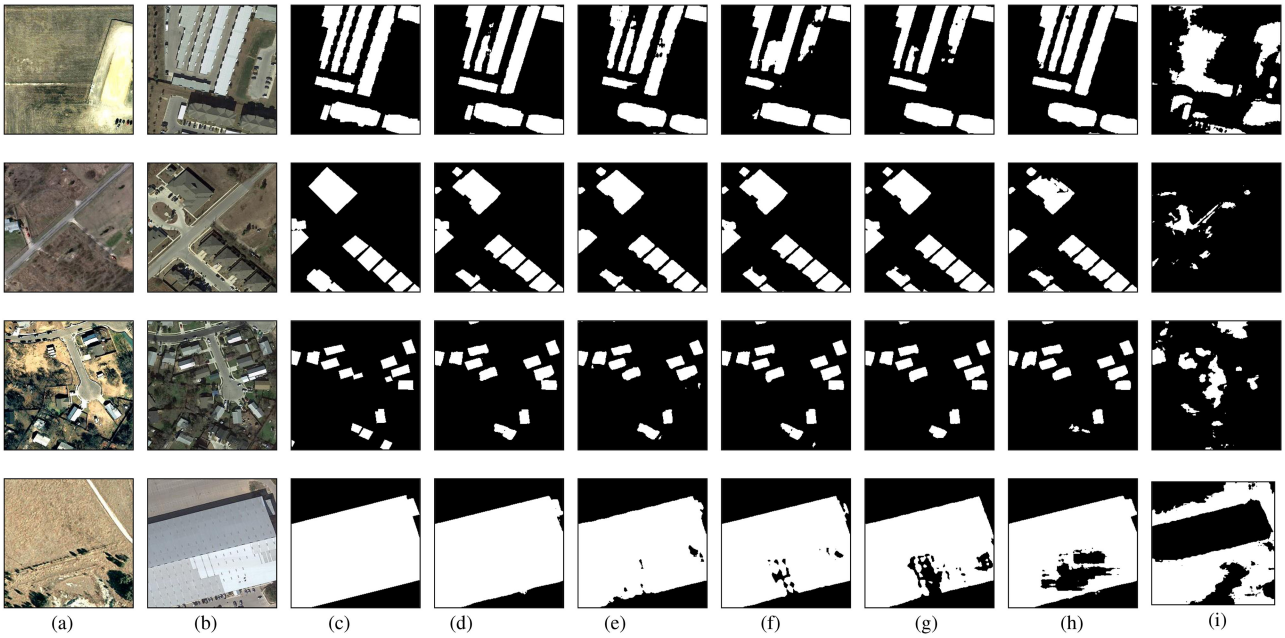

Fig. 11. Ablation study visualization with 256×256 pixel patches. Here, B is base network, D is DSCM, S is SAM, M is MSAM, R is residual connections, and Da is data augmentation at test time. (a) Image 1. (b) Image 2. (c) Groundtruth. (d) B + M + D + R + Da. (e) B + M + D + R. (f) B + M + D. (g) B + M. (h) B + S. (i) B.

pseudochanges. The visual output comparison of the proposed work, FC-EF [11], SNUNet [24], and ground truth is presented in Figs. 7–10. The competitive methods are chosen based on their recent SOTA performance. For example, SNUNet [24] is the recently published one of the SOTA work. FC-EF [11] is chosen as it is one of the foundation works for DL-based CD. It is [11] used as a baseline to show how the proposed model improved the performance over the base performance. In Fig. 7, CDD dataset is used with changes of buildings and roads. While FC-EF [11] network could not make difference between building and road changes, SNUNet [24] generated better results. Our network produced more precise results compared to these. Fig. 8 contains outputs produced on the SYSU dataset with multiple change types. For this dataset, the proposed method is relatively better than others. The output comparison on the LEVIR-CD dataset with building change instances is shown in Fig. 9. The output of FC-EF [11] missed many building change instances and SNUNet [24] detected some false changes. Our proposed work output is close to the actual ground truth and missed minimal changes. We used another building change dataset, WHU-CD, and its visual outputs are shown in Fig. 10. For this dataset, FC-EF [11] detected several false changes and SNUNet [24] also detected some false changes. Our proposed work detected lesser false changes.

VI. Ablation Study

We conducted extensive experiments to find the optimal fusion strategy for the inputs, the optimal number of initial channels, the number of resolution levels in MSAM, and architecture components. We conducted all experiments on the LEVIR-CD test dataset.

TABLE VII
Ablation Study on Fusion Methods

	Precision	Recall	F1 Score
No fusion	89.86	77.90	83.46
Medium	90.99	85.30	88.05
Early	89.38	87.97	88.67
Early + difference	90.67	88.79	89.72

TABLE VIII
Ablation Study on Attention Module

	Parameters	FLOPS	F1 Score
SAM	9.6 K	44.37 M	90.19
MSAM-2	10.17 K	45.32 M	90.25
MSAM-3	10.75 K	45.83 M	90.38
MSAM-4	11.33 K	46.22 M	90.38

Here, SAM is a self-attention module. MSAM-2 is the multiscale attention module with two resolutions of self-attention maps. MSAM-3 is a multiscale attention module with three resolutions of self-attention maps. MSAM-4 is a multiscale attention module with four resolutions of self-attention maps. Parameters are calculated in thousands (K), and FLOPS are counted in millions (M).

A. Ablation for Fusion Strategy

We started with a base model similar to HRNet [45]. It is observed in the literature that handling multitemporal inputs are an important task for CD. For an optimum strategy to combine the input pair, a comprehensive set of experiments has been conducted as tabulated in Table VII. We tried early fusion of input pairs, a medium fusion of input pairs, no fusion of input pairs, and early fusion with the modulus of the input pair difference. The input pair is concatenated in early fusion and passed to

Fig. 12. 1024×1024 pixel output visualization for LEVIR test images. (a) Image 1. (b) Image 2. (c) Ground truth. (d) Our Output.

TABLE IX
ABLATION STUDY

	Precision	Recall	F1 Score	IoU
Base	90.67	88.79	89.72	80.94
Base + Residual connection	90.31	89.79	90.04	81.03
Base +SAM	90.59	89.81	90.19	82.01
Base + MSAM	90.83	89.95	90.38	82.12
Base + DSCM + MSAM	89.67	91.58	90.61	82.21
Base + DSCM + MSAM + Residual connection	92.09	90.85	91.46	84.28
Base + DSCM + MSAM + Residual connection + DA	92.78	90.77	91.77	84.79
Base + DSCM + MSAM + Residual connection + DAE1	92.96	90.83	91.88	84.98
Base + DAE2	90.69	88.82	89.75	80.94
Base + Residual connection + DAE2	90.36	89.81	90.08	81.04
Base + SAM + DAE2	90.59	89.83	90.21	82.01
Base + MSAM + DAE2	90.81	90.07	90.44	82.16
Base + DSCM + MSAM + DAE2	89.81	91.74	90.76	82.43
Base + DSCM + MSAM + Residual connection + DAE2	93.05	90.91	91.97	85.13

Here, DA is data augmentation at test time, DAE1 is data augmentation at test time with an overlapped evaluation with stride 128 pixels, and DAE1 is data augmentation at test time with an overlapped evaluation with a stride of 64 pixels.

the network for feature extraction. In medium fusion, the input pair is given simultaneously to the base model till the middle of it. After that, features from both streams are combined using a concatenation operation. No fusion strategy used concatenated input features till the final class convolution layer. In early fusion with the modulus of the input pair difference approach, concatenated features of input pairs and the modulus of input pair difference are used. The experiment shows that early fusion with the difference approach gives the best results.

B. Ablation for Initial Channels

The selection of an optimal value for the initial number of channels of BN is essential as it affects the computational requirement of the model. We find the number of initial channels for the base model through detailed experiments as shown in Table VI. We tried three different combinations of 32, 48, and 64 as initial channels. From the experiment, it is found that the initial channel of 48 achieves the best results as it produces *F1 score* similar to the 64 channels but with lesser computational cost.

C. Ablation for Attention Mechanism

In Table VIII, we presented *F1 score* produced by attention modules with different resolution inputs. We used input size of $1 \times 48 \times 256 \times 256$ for calculation of computational complexity of attention modules. We started with SAM with input of $\frac{1}{4}$th

of the original resolution and extended it to multiscale SAM. To find the optimal number of different resolution inputs to be used, we started with MSAM-2 where input resolutions of $\frac{1}{4}$th and $\frac{1}{6}$th are utilized. In MSAM-3, input resolutions of $\frac{1}{4}$th, $\frac{1}{6}$th, and $\frac{1}{8}$th are used. MSAM-4 took $\frac{1}{16}$th resolution input additionally. From the experiments, we have found that MSAM with three resolution features are giving better results quantitatively and with respect to the computational complexity.

D. Ablation for Architecture Component

This study is conducted to showcase the importance of each component in the proposed work. The visual comparison of the ablation study for various components of the proposed model is shown in Fig. 11. The rest of the ablation experiments (as shown in Table IX) are done with the base model with the initial optimum channel number obtained by Table VI and early fusion with the difference strategy. This model achieved *F1 score* of 89.72%. After adding the SAM, it increased by 0.47%. We improved the SAM to MSAM with the base model, and this increases the *F1 score* to 90.38% and by 0.66%. We added residual connections in the base model, and this increases the *F1 score* to 90.04% and by 0.32% from the base model. To further improve the performance, we have added the DSCM. First, we combined the DSCM and MSAM with the base and achieved *F1 score* of 90.61% which is 0.89% higher than the base model. Further inclusion of skip connection has improved it to 91.46% which is 1.74% higher than the base model. After applying test time augmentation, this further improved to 91.77 %. In addition, overlapping strides are used in output as utilized in [26]. Output strides of 128 pixels resulted in the *F1 score* of 91.88, and with stride 64, it reached 91.97%. This study shows that skip connection in the base model improved the model's performance. The combination of DSCM and MSAM also significantly enhances the base model's result.

VII. Conclusion

In this article, we presented a novel architecture DRMNet, a multitasking DL model composed of a BN, an MSAM, and a subpixel convolution based deconvolution module. The proposed model can predict change map and image difference in parallel and uses two loss functions, hybrid loss, and MSE loss. Initial feature fusion strategy with modules of features difference is applied in our network, and it has outperformed the recent best published works. A detailed study is presented for justifying the proposed loss functions. An ablation study is also presented to highlight the contributions of the different modules of the proposed architecture. A comprehensive set of experiments reveal that the proposed model has achieved the SOTA results for CDD, SYSU, and LEVIR-CD datasets. We also have set benchmark results for the BCDD dataset for future comparison.

Acknowledgment

This work is part of the Ph.D. work of Mr. Avinash Chouhan under the Department of Computer Science and Engineering, IIT Guwahati. The authors would like to thank Director North Eastern Space Applications Centre for providing computational resources and guidance during this work. The authors would also like to thank Editor, Associate Editor, and Reviewers for their valuable suggestions to improve the manuscript.

References

[1] J. Liu, M. Gong, K. Qin, and P. Zhang, "A deep convolutional coupling network for change detection based on heterogeneous optical and radar images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 3, pp. 545–559, Mar. 2018.

[2] X. Jiang, G. Li, Y. Liu, X.-P. Zhang, and Y. He, "Change detection in heterogeneous optical and SAR remote sensing images via deep homogeneous feature fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1551–1566, Apr. 2020.

[3] 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.

[4] 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.

[5] C. Zhang, S. Wei, S. Ji, and M. Lu, "Detecting large-scale urban land cover changes from very high resolution remote sensing images using CNN-based classification," *ISPRS Int. J. Geo-Inf.*, vol. 8 no. 4, p. 189, 2019.

[6] C. Wu, B. Du, X. Cui, and L. Zhang, "A post-classification change detection method based on iterative slow feature analysis and Bayesian soft fusion," *Remote Sens. Environ.*, vol. 199, pp. 241–255, 2017.

[7] P. Du, X. Wang, D. Chen, S. Liu, C. Lin, and Y. Meng, "An improved change detection approach using tri-temporal logic-verified change vector analysis," *ISPRS J. Photogramm. Remote Sens.*, vol. 161, pp. 278–293, 2020.

[8] C. Wu, B. Du, and L. Zhang, "Slow feature analysis for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2858–2874, May 2014.

[9] J. Im, J. R. Jensen, and J. A. Tullis, "Object-based change detection using correlation image analysis and image segmentation," *Int. J. Remote Sens.*, vol. 29, no. 2, pp. 399–423, 2008.

[10] A. Lefebvre, T. Corpetti, and L. Hubert-Moy, "Object-oriented approach and texture analysis for change detection in very high resolution images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2008, vol. 4, pp. 663–666.

[11] R. C. Daudt, B. L. Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proc. IEEE Int. Conf. Image Process.*, 2018, pp. 4063–4067.

[12] X. Peng, R. Zhong, Z. Li, and Q. Li, "Optical remote sensing image change detection based on attention mechanism and image difference," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7296–7307, Sep. 2021.

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

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

[15] J. Chen *et al.*, "DASNet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1194–1206, Nov. 2021.

[16] M. Papadomanolaki, S. Verma, M. Vakalopoulou, S. Gupta, and K. Karantzalos, "Detecting urban changes with recurrent neural networks from multitemporal sentinel-2 data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 214–217.

[17] H. Du *et al.*, "Bilateral semantic fusion siamese network for change detection from multitemporal optical remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, Jun. 2022.

[18] 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.

[19] 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.

[20] Y. Zhang, S. Zhang, Y. Li, and Y. Zhang, "Coarse-to-fine satellite images change detection framework via boundary-aware attentive network," *Sensors*, vol. 20, no. 23, 2020, Art. no. 6735.
[21] C. Zhang *et al.*, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 183–200, 2020.
[22] D. Wang, X. Chen, M. Jiang, S. Du, B. Xu, and J. Wang, "Ads-net: An attention-based deeply supervised network for remote sensing image change detection," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 101, 2021, Art. no. 102348.
[23] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 811–815, May 2021.
[24] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected siamese network for change detection of VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, Feb. 2022.
[25] 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.
[26] I. Foivos, F. D. Waldner, and P. Caccetta, "Looking for change? Roll the dice and demand attention," *Remote Sens.*, vol. 13, no. 18, 2021, Art. no. 3707.
[27] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, Jul. 2022.
[28] W. Zhao, X. Chen, X. Ge, and J. Chen, "Using adversarial network for multiple change detection in bitemporal remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, Nov. 2022.
[29] H. Chen, W. Li, and Z. Shi, "Adversarial instance augmentation for building change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, Mar. 2022.
[30] M. A. Lebedev, Y. V. Vizilter, O. V. Vygolov, V. A. Knyaz, and A. Y. Rubis, "Change detection in remote sensing images using conditional adversarial networks. ISPRS—International archives of the photogrammetry," *Remote Sens. Spatial Inf. Sci.*, vol. 422, pp. 565–571, May 2018.
[31] M. Liu, Q. Shi, A. Marinoni, D. He, X. Liu, and L. Zhang, "Super-resolution-based change detection network with stacked attention module for images with different resolutions," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, Jul. 2022.
[32] H. Zhang, M. Lin, G. Yang, and L. Zhang, "ESCNet: An end-to-end superpixel-enhanced change detection network for very-high-resolution remote sensing images," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2021.3089332.
[33] X. Zhang *et al.*, "DifUnet++: A satellite images change detection network based on Unet++ and differential pyramid," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, Jan. 2021.
[34] Y. Lin, S. Li, L. Fang, and P. Ghamisi, "Multispectral change detection with bilinear convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1757–1761, Oct. 2020.
[35] L. Mou, L. Bruzzone, and X. X. Zhu, "Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 924–935, Feb. 2019.
[36] J. Qu, Y. Xu, W. Dong, Y. Li, and Q. Du, "Dual-branch difference amplification graph convolutional network for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, Dec. 2022.
[37] J. Qu, S. Hou, W. Dong, Y. Li, and W. Xie, "A multilevel encoder–decoder attention network for change detection in hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, Nov. 2022.
[38] 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.
[39] H. Chen, C. Wu, B. Du, L. Zhang, and L. Wang, "Change detection in multisource VHR images via deep siamese convolutional multiple-layers recurrent neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2848–2864, Apr. 2020.
[40] H. Lyu, H. Lu, and L. Mou, "Learning a transferable change rule from a recurrent neural network for land cover change detection," *Remote Sens.*, vol. 8, no. 6, p. 506, 2016.
[41] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.
[42] L. Ru, B. Du, and C. Wu, "Multi-temporal scene classification and scene change detection with correlation based fusion," *IEEE Trans. Image Process.*, vol. 30, pp. 1382–1394, Nov. 2021.
[43] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, Jun. 2022.
[44] 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.
[45] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.
[46] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," in *Advances in Neural Information Processing Systems*, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, Inc., 2019.
[47] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1874–1883.
[48] H. Lee *et al.*, "Local similarity siamese network for urban land change detection on remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4139–4149, Mar. 2021.
[49] P. Chen, D. Hong, Z. Chen, X. Yang, B. Li, and B. Zhang, "FCCDN: Feature constraint network for VHR image change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 187, pp. 101–119, 2022.

Avinash Chouhan received the B.E. degree in computer science and engineering from Government Engineering College Jabalpur, Jabalpur, India, in 2012. He is currently working toward the Ph.D. degree in deep learning based model development for change detection and semantic segmentation using remote sensing data with the Department of Computer Science and Engineering, IIT Guwahati, Guwahati, India.

He joined North Eastern Space Applications Centre, Umiam, Meghalaya, in 2014 and is currently working as a Scientist/Engineer-SD.

Arijit Sur received the M.Sc. degree in computer and information science and the M.Tech. degree in computer science and engineering from the Department of Computer Science and Engineering, University of Calcutta, Kolkata, India, in 2001 and 2003, respectively, and the Ph.D. degree in computer science and engineering from the Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, Kharagpur, India, in 2010.

He is currently working as an Associate Professor with the Department of Computer Science and Engineering, IIT Guwahati, Guwahati, India.

Dibyajyoti Chutia (Senior Member, IEEE) received the M.Tech. degree in information technology and the Ph.D. degree in computer science & engineering with specialization in soft computing technique for classification of geospatial data from Tezpur Central University, Tezpur, India, in 2000 and 2015, respectively.

He is a Scientist-SF and Team Lead, IT & GeoInformatics, North Eastern Space Applications Centre (NESAC), Department of Space, Umiam, India. His research interests include soft computing, machine learning, and deep learning application for remotely sensed data. He has authored or coauthored more than 65 research articles in peer-reviewed journals, IEEE proceedings, conferences, book chapters, etc.