Remote Sensing Change Detection with Transformers Trained from Scratch

Mubashir Noman*, Mustansar Fiaz*, Hisham Cholakkal, Sanath Narayan, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan

***Abstract*—Current transformer-based change detection (CD) approaches either employ a pre-trained model trained on large-scale image classification ImageNet dataset or rely on first pre-training on another CD dataset and then fine-tuning on the target benchmark. This current strategy is driven by the fact that transformers typically require a large amount of training data to learn inductive biases, which is insufficient in standard CD datasets due to their small size. We develop an end-to-end CD approach with transformers that is trained from scratch and yet achieves state-of-the-art performance on five benchmarks. Instead of using conventional self-attention that struggles to capture inductive biases when trained from scratch, our architecture utilizes a shuffled sparse-attention operation that focuses on selected sparse informative regions to capture the inherent characteristics of the CD data. Moreover, we introduce a change-enhanced feature fusion (CEFF) module to fuse the features from input image pairs by performing a per-channel re-weighting. Our CEFF module aids in enhancing the relevant semantic changes while suppressing the noisy ones. Extensive experiments on five CD datasets reveal the merits of the proposed contributions, achieving gains as high as 1.35% in intersection over union (IoU) score, compared to the best-published results in the literature. Code is available at https://github.com/mustansarfiaz/ScratchFormer.**

***Index Terms*—Remote Sensing, Change Detection, Transformers.**

I. Introduction

CHANGE DETECTION (CD) is a fundamental remote sensing research problem that strives to identify all relevant changes between co-registered satellite images acquired at distinct timestamps. CD plays a crucial role in various remote sensing applications including, disaster management [1], urban planning [2], forestry and ecosystem monitoring [3], [4]. The objective of the CD task is to detect relevant semantic changes in man-made facilities such as buildings and other constructions while ignoring noisy changes such as shadows, illumination variations, and all types of seasonal and environmental variations. Fig. 1 shows a few challenges related to the CD problem in bi-temporal satellite images. For instance, trees, shadows, and cars in Fig. 1-(a) and (b) may limit the detection performance. Likewise, illumination variations and environmental conditions may affect the color of the objects as depicted in Fig. 1-(c). In Fig. 1-(d), accurate detection of subtle and large change regions is a challenging task due to scale variations and densely constructed regions. Lastly, Fig. 1-(e), highlights the weather condition challenges which may affect the CD performance. Therefore, extracting meaningful feature representations while neglecting the irrelevant information is necessary for the CD task.

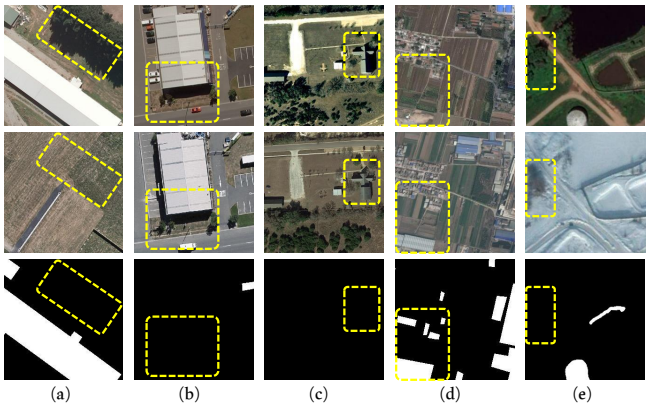

Fig. 1. Figure illustrates various challenges in satellite images for the change detection task including (a) trees and their shadows, (b) cars and shadow directions, (c) illumination variations, (d) scattered subtle and large semantic change regions, and (e) seasonal conditions. The first, second, and third rows indicate the pre-change, post-change, and ground truth images, respectively.

Generally, CD approaches relying on convolutional neural networks (CNN) have shown promising results by utilizing explicit mechanisms such as dilated convolutions, channel and spatial attentions. Zhang et al. [5] utilized atrous convolution based CNN to obtain dense feature representation. DASNet [6] uses dilated convolutions along with the standard convolutions to extract local feature representations and apply dual attention mechanism to further enhance those features. Some approaches [7], [8] utilized standard convolutions with pooling layers to extract deep features at multiscale levels. STANet [9] uses standard CNN networks for deep feature extraction and utilizes spatial and temporal attention modules to refine those features representations. However, these CNN-based approaches typically struggle to capture long-range dependencies between different image regions, hampering the change detection performance.

Recently, transformer-based CD methods [10], [11], [12], [13], [14], [15], [16] have shown competitive performance on various CD datasets by capturing long-range dependencies be-

Mubashir Noman and Mustansar Fiaz have equal contributions (e-mail: mubashir.noman@mbzuai.ac.ae, mustansar.fiaz@ibm.com).

Mubashir Noman and Hisham Cholakkal are with the Mohamed bin Zayed University of Artificial Intelligence, UAE. Mubashir Noman is also the corresponding author (Email: mubashir.noman@mbzuai.ac.ae).

Mustansar Fiaz is with IBM Research.

Sanath Narayan is with the Technology Innovation Institute, UAE.

Rao M. Anwer is with the Mohamed bin Zayed University of Artificial Intelligence, UAE and Aalto University, Finland.

Salman Khan is with the Mohamed bin Zayed University of Artificial Intelligence, UAE and Australian National University, Australia

Fahad S. Khan is with the Mohamed bin Zayed University of Artificial Intelligence, UAE and Linköping University, Sweden.

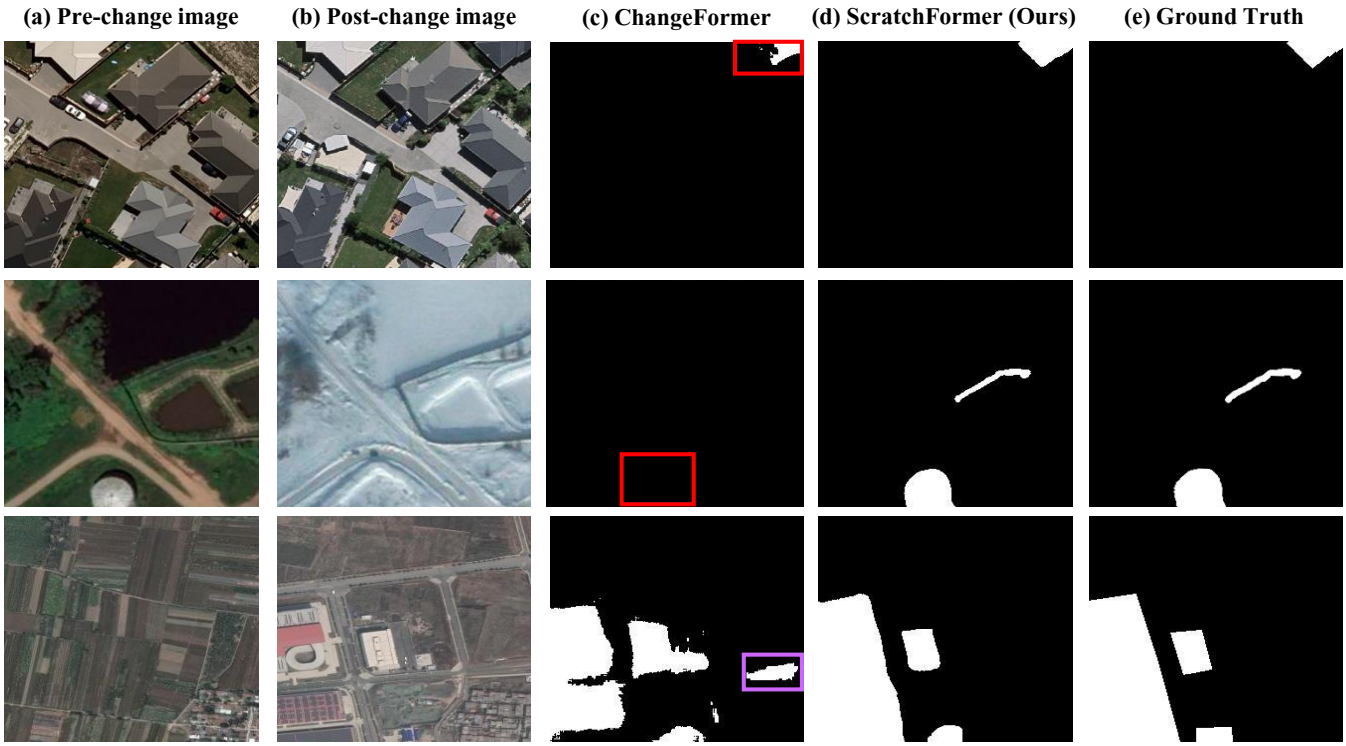

Fig. 2. Change detection performance comparison of **(d)** our approach (ScratchFormer) with **(c)** the recent ChangeFormer. Here, **(a)** the pre-change image and **(b)** post-change image are shown along with **(e)** the ground-truth. We show the false positives and false negatives in the purple and red colors, respectively. In the first two rows, the ChangeFormer fails to detect the change occurring between the pre- and post-images (red box in both rows). Similarly, the ChangeFormer incorrectly detects a change region (purple box), as indicated in the ground-truth. Our ScratchFormer achieves improved change detection performance, in different challenging scenarios, by reducing both false positives and negatives. Best viewed zoomed in.

tween uniformly sampled dense patches through self-attention [17]. Although achieving superior CD performance, state-of-the-art transformer-based methods [10] generally require *pre-training* based weight initialization for optimal convergence. The pre-training step in existing transformer-based CD methods either involves another CD dataset [10] or an ImageNet pre-trained image classification model [16], [11], [12], [18], [14]. In addition, several Siamese network-based supervised contrastive pretraining methods have been proposed to handle overfitting, but random initialization lacks any prior CD knowledge [19], [20]. Zhang et al. [21] studied different metric learning and proposed spatial–temporal triplet loss (STTL) for the CD task. However, the performance of these transformer-based CD methods drastically reduces when directly training from scratch on the target CD dataset. This is likely due to the dense self-attention operation, utilized in these approaches, which has quadratic complexity with respect to tokens, requires longer to converge, and is prone to overfitting. In this work, we look into the problem of designing a transformer-based CD approach that is capable of achieving high performance when trained from *scratch*.

Most existing transformer-based CD approaches employ a two-stream architecture, where features from both streams are combined through simple operations such as difference, summation, and concatenation [22], [10]. However, these approaches do not employ any explicit feature re-weighting between both streams. We argue that such naive feature fusion strategies likely struggle to effectively aggregate semantic changes from each stream. In this work, we set out to address the above issues collectively in a single transformer-based CD architecture.

In this paper, we propose a transformer-based Siamese two-stream CD framework, named ScratchFormer, that is based on a novel shuffled sparse attention (SSA) operation that strives to better attend to sparse informative regions relevant to the CD task. The proposed SSA performs token-mixing over a sparse subset of shuffled features obtained through a data-dependent feature sampling, enabling optimal CD performance when being trained from scratch directly on the target CD dataset. Furthermore, we introduce a change-enhanced feature fusion module (CEFF) that performs feature fusion based on per-channel re-calibration to enhance the features relevant to the semantic changes, while suppressing the noisy ones. We perform extensive experiments on five public CD datasets: LEVIR-CD [9], DSIFN-CD [7], WHU-CD [23], CDD-CD [24], and OSCD [25]. Our proposed ScratchFormer approach achieves superior performance over the baseline, highlighting the effectiveness of the proposed contributions. Compared to the baseline, our ScratchFormer achieves an absolute gain of 1.35% in terms of intersection over union (IoU) on the CDD-CD dataset. Furthermore, ScratchFormer sets a new state-of-the-art performance on all five datasets. On the LEVIR-CD, our ScratchFormer achieves an IoU score of 84.63%, outperforming the recent method [10] published in literature

by 2.15%. Fig. 2 shows a qualitative comparison between the recent ChangeFormer [10] and our ScratchFormer on different challenging CD examples.

We summarize our contributions as follows:

- We propose a hierarchical transformer-based Siamese two-stream change detection algorithm, dubbed as ScratchFormer, which is trained from scratch and yet achieves state-of-the-art performance, hence removing the pre-training requirement on another CD dataset and then fine-tuning on the target benchmark.
- We propose a shuffled sparse attention that benefits from the sparse informative regions for the CD task. The proposed SSA operation performs token-mixing over a sparse subset of shuffled features obtained through data-dependent feature sampling.
- We propose a change-enhanced feature fusion module that is responsible for highlighting the semantic features while ignoring the noisy ones.
- Extensive experiments on five CD datasets validate the merits of our proposed algorithm. Our algorithm shows state-of-the-art performance compared to CNN, transformer, and hybrid CD approaches.

II. Related Work

Convolutional neural networks have attained much popularity in remote sensing change detection due to intrinsic properties to capture discriminative features [18]. Chen et al. [6] propose a dual attention mechanism within Siamese CNN to encode long-range dependencies. The Siamese CNN module is used to extract local features from the image pairs. Then, a dual attention module is utilized to obtain the global contextual features for better separation of changed and unchanged regions. Fang et al. [26] propose a dense Siamese network to extract features from bi-temporal images and use an ensemble channel attention module to refine and aggregate the features at multiple semantic levels. The aim of the proposed module is to suppress the semantic gaps and reduce the localization error of the change regions. A feature pyramid with attention mechanism is proposed to encode long-range dependencies in [27]. Authors utilize VGG16 [28] network as a backbone feature extractor. A co -attention module is then used to aggregate the low and high-level features for better detection results. Liu et al. [29] use multi-scale convolutional attention features to learn the bi-temporal feature differences via adversarial learning. The authors employ a super-resolution module consisting of a generator and a discriminator to learn the mapping between a low and high-resolution image via adversarial learning. Then a stacked attention module is utilized to enhance the discriminative features at multi-scale level. Hou et al. [30] employ low rank analysis to benefit from deep features for CD. Xu et al. [31] propose MFPNet which performs channel attention for the CD task. Similarly, Wang et al. [32] make use of spatial and channel attention to improve feature representations. RaSRNet [33] introduces a graph-based relation-aware to handle the restricted receptive field of CNNs for the CD task. Chen et al. [9] introduce Siamese-based network to capture spatial–temporal dependencies using spatial attention and channel attention. Zhang et al. [7] propose a deep supervised image fusion network for CD. A Siamese-based CNN is used for feature extraction from bi-temporal images. The extracted features are input to the difference discrimination network and the change detection mask is obtained through deep supervision.

Recently, transformers [17] have gained popularity for the CD task [34]. Chen et al. [16] introduce a bitemporal image transformer (BIT) to model context information. BIT utilizes ResNet18 [35] for feature extraction from the remote sensing image pairs. The extracted features are converted to a set of semantic tokens and a contextual relationship is modeled between the sets of token features through a transformer encoder. The encoded features are projected back to spatial space by utilizing a Siamese transformer decoder. A shallow CNN module is then used to predict the change mask. Li et al. [11] introduce TransUNetCD, which benefits from both transformers and UNet for CD. TransUNetCD utilizes a CNN to extract features from bi-temporal images followed by a transformer to obtain better discriminative features for the change detection task. Zhou et al. [36] use self-attention to model contextual-semantic relations between the input bi-temporal images. Zhao et al. [37] propose a position matching mechanism (PMM) to perform sparse pixel-level adaptive matching of multitemporal images utilizing geospatial position and content reasoning mechanism (CRM) to discriminate the diverse pseudo-change information. Hu et al. [38] employ an unsupervised joint learning model utilizing total variation regularization and bipartite CNN. Fang et al. [39] propose a Siamese network to compute multi-layered features and perform feature exchange operations for the two steams for the CD task.

Zhang et al. [40] utilize hierarchical Swin transformer [41] to extract global information in bi-temporal images. Song et al. [12] utilizes a multi-scale Swin transformer to enhance the extracted features from a Siamese-based CNN network at the multi-scale level. Wang et al. [42] propose STCN which exploits cross Swin transformer to extract global features. Teng et al. [43] introduce SFCD which improves representation using the foreground aware fusion module to use attention gates to trim low-level feature responses. Hong et al. [44] integrate the multi-task network into Swin transformer to use the available training samples for representative feature learning. Ke et al. [14] propose a hybrid transformer to capture global context dependencies at multiple scales. After extracting features from a CNN backbone, proposed hybrid transformer is utilized to learn the global context relationships before input to the cascade decoder for change map prediction. Bandara et al. [10] propose a hierarchical Siamese transformer to render multi-scale features. The Siamese encoder utilizes self attention and a convolution layer to learn the discriminative features. Different to existing approaches, we introduce SSA to effectively capture the inductive CD bias when training from scratch on any change detection (target) dataset. Further, a CEFF module to perform per-channel re-weighting to enhance the feature channels having higher semantic changes, while suppressing the channels encoding noisy changes.

III. Preliminaries

Problem Formulation: Given I_{pre}, $I_{post} \in \mathbb{R}^{3\times H\times W}$ as a pair of co-registered satellite images acquired at distinct times T_1 and T_2, the objective in CD is to detect relevant semantic changes between I_{pre} and I_{post} while ignoring irrelevant changes. Here, the relevant changes include changes in man-made facilities such as buildings and other constructions. On the other hand, the irrelevant changes include seasonal variations, illumination changes, building shadows, and atmospheric variations. Consequently, the goal in CD is to predict a binary mask $\boldsymbol{M} \in \mathbb{R}^{H\times W}$ that depicts the semantic (structural) changes between I_{pre} and I_{post}.

A. Baseline Change Detection Framework

We adapt the recently introduced transformer-based approach [10] as our base framework since it achieves promising performance for the CD task. The base CD framework takes an image pair as input and computes the semantic difference between them using a transformer-based Siamese network. It comprises a transformer encoder, difference feature fusion module, and a decoder. The encoder consists of a series of attention layers with each layer comprising the standard self-attention [17] followed by a feed-forward network. The encoder weights are shared and utilized for computing multi-scale features in both streams (pre-change and post-change). For each scale i, the resulting features F^i_{pre}, F^i_{post} from both the streams are input to a difference feature fusion module, which encodes the semantic changes occurring between the streams in the corresponding scale. The difference feature fusion module comprises a feature concatenation followed by two convolutions with batch normalization and ReLU layers in between. It then outputs the feature F^i_{diff} for scale i. These multi-scale features F^i_{diff} are then input to the decoder, where they are fused through a series of convolution and transpose convolution layers for increasing the spatial resolution of feature maps. Finally, the resulting upsampled features are passed to a mask prediction layer to obtain final semantic binary change map $\boldsymbol{M}$.

Limitations: As discussed above, the base framework employs the transformer encoder with the standard self-attention mechanism to capture long-range dependencies in an image. Here, we argue that the standard self-attention mechanism is sub-optimal for the CD task mainly due to the following reasons. It operates on uniformly sampled dense patches, thereby requiring large training data for optimal convergence in terms of CD performance (see Fig. 3). The recent ChangeFormer [10] alleviates this issue by performing pre-training on one (source) CD dataset followed by fine-tuning on another (target) CD data. However, this increases the training time when including the cost of pre-training on another CD dataset as well. Furthermore, despite being trained from scratch the proposed ScratchFormer outperforms our baseline accuracy that is achieved through a pre-training step on another CD dataset, with less than 50% of the training time.

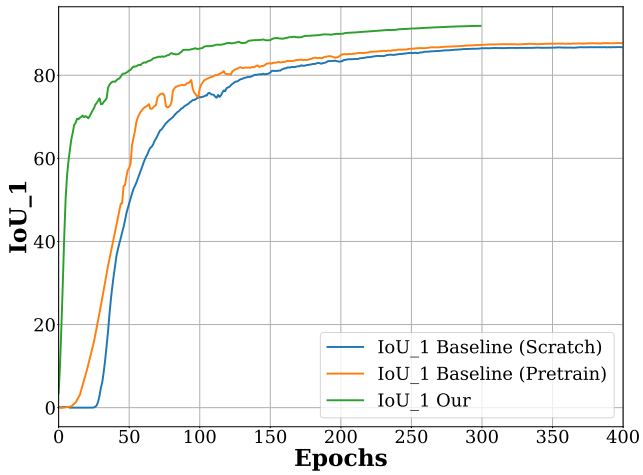

Fig. 3. Comparison, in terms of intersection over union (IoU) vs. the training epochs, among the baseline trained from scratch, baseline pre-trained† first on another CD dataset and then fine-tuned, and our approach on the LEVIR-CD. Compared to the baseline employing pre-training, training the baseline from scratch results in inferior convergence in terms of CD performance. Our approach despite being trained from scratch achieves superior convergence in terms of CD performance compared to both variants of baseline. For instance, with less than 50% of the training time, our approach achieves similar CD performance to that of the final results obtained from the baseline trained from scratch.

IV. Method

A. Motivation

To motivate our proposed approach, we distinguish two properties especially desired when designing a transformer-based CD method.

1) Rethinking Attention for CD Task: As discussed earlier, the conventional self-attention may lead to sub-optimal performance when training from scratch directly on the target CD dataset, likely due to difficulty in capturing the inherent inductive biases in the small CD dataset. Moreover, the standard self-attention typically operates on uniformly sampled dense patches that may have difficulties to learn a rich feature representation encoding diverse shape objects with inconsistent appearance in remote sensing scenes having sparse informative regions. Therefore, rethinking the design of self-attention is desired to effectively learn a rich feature representation by attending to sparse informative regions in remote sensing CD images.

2) Semantic Change-enhanced Feature Fusion: Though the above requisite focuses on designing a mechanism to attend sparse informative regions for the CD task, the second desirable characteristic aims at capturing the semantic differences between image pairs while ignoring the irrelevant noisy changes. To this end, a change-enhanced feature fusion module that explicitly models per-channel inter-dependencies between pre- and post-change images is expected to better ignore the noisy changes while retaining the relevant ones. Next, we present our proposed transformer-based ScratchFormer framework.

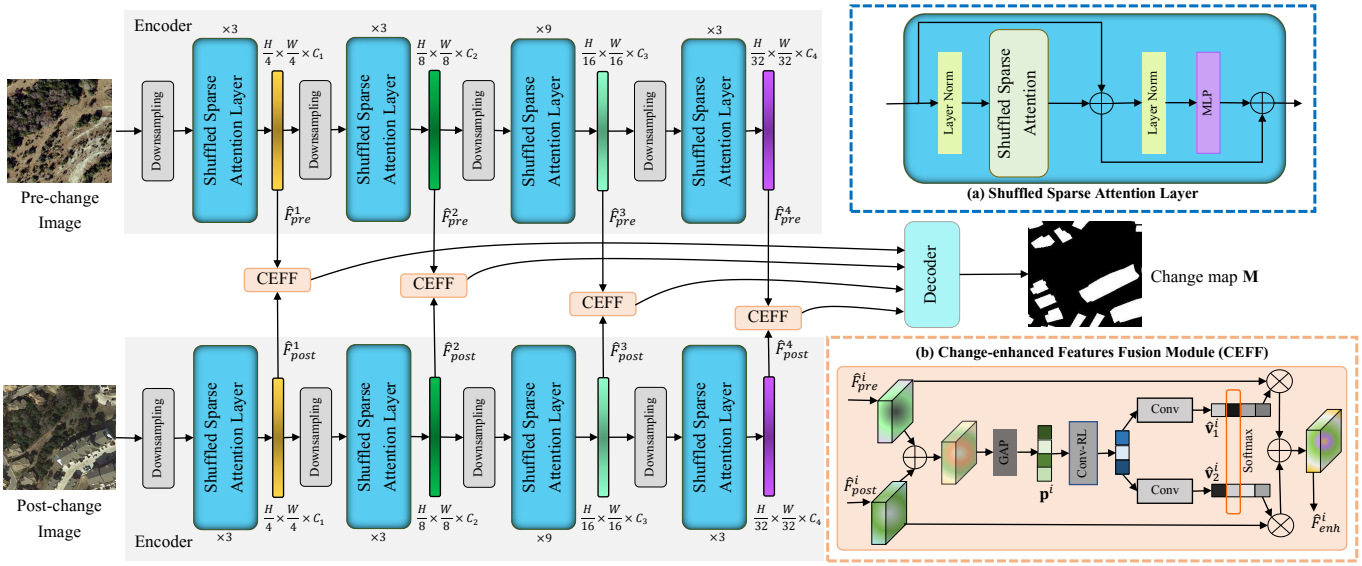

Fig. 4. Overall architecture of our **ScratchFormer** framework for Change Detection. Our ScratchFormer takes two inputs, pre- and post-change images, and predicts a binary semantic change map for the corresponding image pair. ScratchFormer consists of a Siamese-based hierarchical encoder having four different stages, a change-enhanced feature fusion (CEFF) module, and a decoder for predicting binary change map. The focus of our design is the introduction of a *shuffled sparse attention* (SSA) layer (Sec. IV-C) in the encoder and a *change-enhanced feature fusion* (CEFF) module (Sec. IV-D). The SSA layer comprises shuffled sparse attention and a MLP, as shown in (a). SSA performs token-mixing over a sparse data-dependent subset of features at each stage. Our ScratchFormer approach computes SSA features from the two streams $\hat{F}^i_{pre}$ and $\hat{F}^i_{post}$ at different scales i. The outputs of these stages are fused utilizing the CEFF module, as shown in (b). The CEFF module enhances the semantic changes between the features of the two streams by performing a per-channel re-weighting at each scale and outputs enhanced features $\hat{F}^i_{enh}$. These enhanced features are then input to the decoder for predicting the final semantic binary change map $\boldsymbol{M}$.

B. Overall Architecture

Fig. 4 shows the overall architecture of our ScratchFormer. The proposed ScratchFormer takes pre- and post-change image pairs (I_{pre}, I_{post}) as input. It comprises a Siamese-based encoder, a change-enhanced feature fusion module, and a decoder for predicting the binary change map $\boldsymbol{M}$. The encoder computes the features at four stages with different spatial resolutions. At each stage, the features are first spatially downsampled through convolutional layers and then input to the SSA layers. The ScratchFormer consists of two parallel identical encoder streams with shared weights to generate pre- and post- change features $\hat{F}^i_{pre}$, $\hat{F}^i_{post}$, respectively at the i-th stage of our multi-stage network. The focus of our design is the introduction of a novel *shuffled sparse attention* layer in the encoder to perform the self-attention on the data-dependent subset of features to effectively capture the semantic changes for CD task. Furthermore, we propose a *change-enhanced feature fusion* module that re-calibrates the per-channel features of the same scale from both streams ($\hat{F}^i_{pre}$ and $\hat{F}^i_{post}$) and performs enhanced feature fusion to better ignore the noisy changes while retaining the relevant ones.

Our SSA layer comprises a shuffled sparse attention and a multi-layer-perceptron (MLP), as shown in Fig. 4-(a). SSA first performs a data-dependent sampling of features to obtain a subset and then performs token-mixing over the selected subset. SSA strives to focus on the sparse informative regions for change detection to achieve optimal convergence with respect to CD performance without requiring pre-training on another CD data. The CEFF modules aims to enhance the semantic changes between pre- and post- change features at each stage of the encoder, while suppressing the noisy changes. The resulting enhanced features from the CEFF module are re-sized to a common spatial resolution and passed to the decoder. The decoder has a series of convolution, transpose convolution, and upsampling layers to increase the spatial resolution of the feature maps. Consequently, these upsampled features are passed to a mask prediction layer to obtain the final binary mask $\boldsymbol{M}$. We also present Algorithm 1 for a better understanding of our overall framework. Next, we present our SSA layer.

C. Shuffled Sparse Attention Layer

We introduce a shuffled sparse attention layer within our encoder to capture semantic changes between the input image pairs I_{pre} and I_{post}.

As shown in Fig. 4-(a), it comprises a shuffled sparse attention to perform token-mixing, a multi-layer perceptrons, and layer normalization layers. Our SSA performs token-mixing over a sparse subset of features which are selected based on a data-dependent sampling strategy. Let, $F^i \in \mathbb{R}^{C^i \times H^i \times W^i}$ be the encoder feature at stage i input to SSA. Then, our SSA is computed in two steps. First, we perform a data-dependant sparse sub-sampling of input features with a sparsity factor of γ to obtain feature sub-sets $\bar{F}^i_{kl}$. Then, we

separately perform self-attention over these γ^2 feature subsets $\bar{F}^i_{kl}$, where $k = \{0, ..., \gamma-1\}$ and $l = \{0, ..., \gamma-1\}$. The data-dependant sparse spatial sub-sampling of features is performed as follows:

$$\bar{F}^i_{kl}(\bar{x}, \bar{y}) = F^i(\gamma\bar{x} + k + \Delta x,\ \gamma\bar{y} + l + \Delta y)$$
$$\forall \bar{x} = \{0, ..., \frac{H^i}{\gamma} - 1\} \text{ and } \forall \bar{y} = \{0, ..., \frac{W^i}{\gamma} - 1\}. \quad (1)$$

Here, $(\Delta x,\ \Delta y)$, represents the data-dependent position offsets which are predicted using learnable parameters θ_{offset} as $\Delta z = \theta_{offset}(F^i)$ [45]. The predicted offsets $\Delta z \in \mathbb{R}^{2\times H^i \times W^i}$ have two channels depicting the horizontal and vertical position offsets at each pixel, which are clipped to limit the maximum distance from the current feature location. Then, the position offsets Δx, Δy are obtained as:

$$\Delta x = \Delta z(\gamma\bar{x} + k, \gamma\bar{y} + l, 1)$$
$$\Delta y = \Delta z(\gamma\bar{x} + k, \gamma\bar{y} + l, 2).$$

The resulting sparse-sampled features $\bar{F}^i_{kl}$ are used to compute self-attention [17] ($Attention(.)$) over the γ^2 sparse windows as follows:

$$\hat{F}^i_{kl} = Attention(\bar{F}^i_{kl}) \quad (2)$$

These attended features $\hat{F}^i_{kl}$ from γ^2 feature subsets are then shuffled back to the original resolution feature map to obtain $\hat{F}^i \in \mathbb{R}^{C^i\times H^i\times W^i}$. Here, the data-dependent position offsets aid in adaptively sampling dense features from regions likely to have semantic changes, whereas the sparse sampling helps to efficiently maintain the global receptive field. For better understanding, we present the flowchart for the computation of data-dependent shuffled-sparsed feature samples in Fig. 5. Due to the sparse sampling, we perform γ^2 self-attention operations and in each self-attention operation the number of tokens are reduced by a factor of γ^2, leading to a $O(\gamma^2)$ reduction in the overall computation. Our SSA enables faster convergence due to its sparse structure allowing self attention to focus on the sub-sampled relevant features. Our proposed ScratchFormer approach employs SSA layers at each stage of the encoder and computes pre- and post- change features $\hat{F}^i_{pre}$, $\hat{F}^i_{post}$, from both streams of the encoder. These features are then fused by the change-enhanced feature fusion module described next.

D. Change Enhanced Features Fusion Module

As discussed earlier, given the diverse nature of the changes in real-world scenarios that can possibly occur in the image pairs, detecting high-level semantic changes while ignoring the noisy ones is one of the major challenges in the CD task. Therefore, it is desired to effectively fuse the features from pre- and post-change feature streams of the encoder. Within several existing transformers-based CD methods [16], [10], [46], multi-level feature fusion between pre- and post change features is performed through difference, summation or concatenation operations. Similarly, the base framework also introduces a difference module employing concatenation across channel dimension for the feature fusion. We argue that such a fusion of the features from both streams without explicitly re-weighting the channels from each stage is sub-optimal for the CD task. To this end, we introduce a CEFF

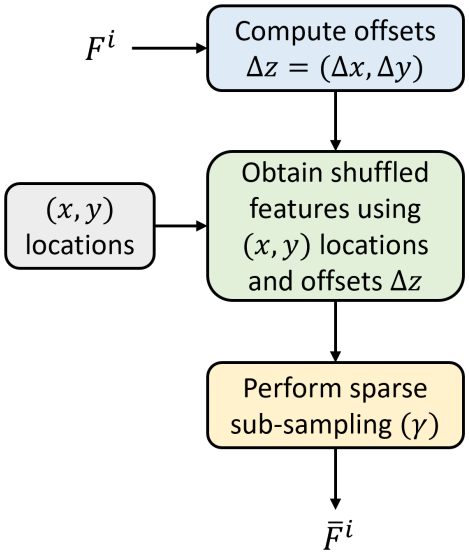

Fig. 5. Illustration of sparse feature shuffling depicted in Eq. 1. The input features F^i are passed to offset computation layer to generate the data-dependent position offsets Δz. Sparse sub-sampled features are extracted using sparsity factor γ. The computed offsets are then utilized to get shuffled sparse features $\bar{F}^i$.

module that performs per-channel re-weighting to enhance the channels having higher semantic changes, while suppressing the channels capturing noisy changes. Fig 4-(b) shows the structure of our change-enhanced feature fusion module. The CEFF module is introduced at all four stages of the encoder to fuse the features at each stage.

In our CEFF module, we first combine the pre- and post-change features $\hat{F}^i_{pre}$, $\hat{F}^i_{post}$ through addition, and then perform global average pooling (GAP) to obtain a global feature vector $\mathbf{p}^i$ as follows:

$$\mathbf{p}^i = GAP(\hat{F}^i_{pre} + \hat{F}^i_{post}), \quad (3)$$

We input $\mathbf{p}^i$ feature vector to shared Conv-ReLU layers to reduce the number of channels. Afterwards, these reduced features are passed to separate 1×1 conv layers to obtain the channel weights for both streams $\mathbf{v}^i_1$, $\mathbf{v}^i_2$ as follows:

$$\bar{\mathbf{p}}^i = \varphi(\omega_1(\mathbf{p}^i)),$$
$$\mathbf{v}^i_1 = \omega_2(\bar{\mathbf{p}}^i),\ \mathbf{v}^i_2 = \omega_3(\bar{\mathbf{p}}^i), \quad (4)$$

where, ω_1, ω_2, and ω_3 are the convolutional weights, and φ represents the ReLU activation function. Here, $\mathbf{v}^i_1 \in \mathbb{R}^{C^i\times 1}$, and $\mathbf{v}^i_2 \in \mathbb{R}^{C^i\times 1}$ refers to the un-normalized channels re-weighting factors predicted for the pre- and post-change features at stage i. These un-normalized weights are then normalized by per-channel softmax across both streams. i.e,

$$\hat{\mathbf{v}}^i_1(j) = \frac{\exp(\mathbf{v}^i_1(j))}{\exp(\mathbf{v}^i_1(j)) + \exp(\mathbf{v}^i_2(j))}$$
$$\hat{\mathbf{v}}^i_2(j) = \frac{\exp(\mathbf{v}^i_2(j))}{\exp(\mathbf{v}^i_1(j)) + \exp(\mathbf{v}^i_2(j))} \quad (5)$$
$$\forall j = \{1, ..., C^i\}$$

where, j is the channel index and exp denotes the exponential function. These normalized weights $\hat{\mathbf{v}}_1^i \in \mathbb{R}^{C^i \times 1}$ and $\hat{\mathbf{v}}_2^i \in \mathbb{R}^{C^i \times 1}$ are used to perform channel re-weighting of $\hat{F}_{pre}^i$ and $\hat{F}_{post}^i$ followed by feature fusion through addition to generate the enhanced features $\hat{F}_{enh}^i$ as:

$$\hat{F}_{enh}^i = \hat{\mathbf{v}}_1^i \hat{F}_{pre}^i + \hat{\mathbf{v}}_2^i \hat{F}_{post}^i, \tag{6}$$

The resulting enhanced features from the CEFF module at all stages are then resized to a fixed spatial resolution and passed to the decoder that performs feature upsampling and change map prediction.

Algorithm 1: Proposed algorithm to compute the change map between the two inputs, pre- and post-change images.

Data: Two co-registered images I_{pre} and I_{post}
Result: Change map $\mathcal{M}$
$S \leftarrow stages = 4$;
$B \leftarrow blocks = [3, 3, 9, 3]$
for $j \in (pre, post)$ **do**
  for $i \leftarrow 1$ *to* S **do**
    if $i == 1$ **then**
      $I_j^i \leftarrow I_j$ # Input at stage 1
    else
      $I_j^i \leftarrow \hat{F}_j^{i-1}$ # Input at stage > 1
    $t \leftarrow$ Downsampling(I_j^i)
    repeat
      # shuffled sparse attention (SSA)
      $\Delta z \leftarrow$ Calculate offsets for t # get offsets
      $\bar{t} \leftarrow$ Get sparse sampled features using Δz (Eq. 1)
      $\hat{t} \leftarrow$ Compute self-attention on $\bar{t}$ (Eq. 2)
      $t \leftarrow \hat{t}$
    until B^i;
    $\hat{F}_j^i \leftarrow t$
for $i \leftarrow 1$ *to* S **do**
  # Change enhanced features fusion (CEFF)
  $p^i \leftarrow$ Compute global vector from $(\hat{F}_{pre}^i, \hat{F}_{post}^i)$ using Eq. 3
  $(\hat{\mathbf{v}}_1^i, \hat{\mathbf{v}}_2^i) \leftarrow$ Get normalized weights using p^i (Eq. 4 and 5)
  $\hat{F}_{enh}^i \leftarrow$ Compute enhanced features using Eq. 6
$F^{all} \leftarrow Concatenate(\hat{F}_{enh}^1, \hat{F}_{enh}^2, \hat{F}_{enh}^3, \hat{F}_{enh}^4)$
$\mathcal{M} \leftarrow$ Decoder(F^{all})

V. Experiments

A. Experimental Setup

Datasets: The large-scale *LEVIR-CD [9]:* dataset is for building change detection. It contains 637 high-resolution (0.5m per pixel) image pairs taken from Google Earth with the size of 1024x1024. In our experiments, we use the non-overlapping cropped patches of 256x256, having default data split of train, validation, and test equal to 7120, 1024, and 2048, respectively. The *DSIFN-CD [7]:* dataset is for binary change detection and contains six high-resolution (2m) satellite image pairs from six cities in China. We used the cropped version of the dataset having image size of 256x256 resulting in train, validation, and test sets of size 14400, 1360, and 28 image pairs, respectively. The *CDD-CD [24]:* dataset comprises 11 seasonal varying image pairs including, 7 image pairs of size 4725x2700 pixels and 4 image pairs of size 1900x1000. The image pairs are clipped into 256x256 with data split of 10000, 3000, and 3000 for train, validation, and test set, respectively. The *WHU-CD [23]:* dataset is for building-related change detection and consists of one high-resolution (0.075 m) image pair of size 32507x15354 pixels. This aerial dataset contains a variety of building architectures of different sizes and colors. The dataset is also available with image pairs of size 256x256 pixels having non-overlapping regions and data split of 5947, 743, and 744 image pairs for train, validation, and test sets, respectively. *OSCD [25]:* is a public change detection dataset focusing on urban changes. It comprises of 24 image pairs of Sentinel-2 multi-spectral data taken from the satellite and also available in RGB format. These image pairs belong to different locations in the world. The dataset focuses on construction related changes and the resolution of image pairs is between 10m to 30m. We crop the RGB images to 256x256 size and use the random rotation, and flipping augmentations to increase the size of the dataset.

Evaluation Protocol: Following [10], we evaluate change detection results in terms of *change class* F1-score, *change class* Intersection over Union (IoU) and overall accuracy (OA) on all the datasets. Among these evaluation metrics, the *change class* IoU is the most challenging metric for the CD task.

Implementation Details: Our ScratchFormer takes a pair of images of size $256 \times 256 \times 3$ and computes the features for the two streams at four stages (having 3, 3, 9, and 3 SSA layers), which outputs the features with 64, 128, 320, and 512 channels, respectively. In the proposed SSA, the sparsity factor is calculated as $\gamma = 2^n$, where $n > 0$. The model is trained using pixel-wise cross-entropy loss function. During training, we employ standard data augmentations including, random scale crop, Gaussian blur, random flip, and random color jitter. We train our network using random initialization on 4 NVIDIA A100 GPUs. Following [10], we use the AdamW optimizer with a weight decay 0.01 and beta values equal to (0.9, 0.999). We set the batch size 16, initial learning rate to 4.1e-4, and train for 300 epochs. In our experiments, we used linear decay to decrease the learning rate till the last epoch. The binary change mask $\boldsymbol{M}$ is computed using a pixel-wise argmax operation along the channel dimension.

B. State-of-the-art Comparison

Comparison on LEVIR-CD: Here, we present the state-of-the-art comparison on the LEVIR-CD dataset (Tab. I). Among recent transformer-based CD methods, H-TransCD [14], BIT [16], ChangeFormer [10], and TransUNetCD [11] obtain IoU scores of 81.92%, 80.68%, 82.48%, and 83.67%, respectively. Our ScratchFormer obtains an IoU score of 84.63% with an absolute gain of 2.15% and 0.96% over the recently published

TABLE I
STATE-OF-THE-ART COMPARISON ON LEVIR-CD, WHU-CD, AND CDD-CD DATASETS. WE REPORT THE RESULTS IN TERMS OF F1, IOU, AND OA METRICS. SCRATCHFORMER PERFORMS SIGNIFICANTLY BETTER AGAINST EXISTING METHODS AND ACHIEVES STATE-OF-THE-ART PERFORMANCE. THE BEST TWO RESULTS ARE IN RED AND BLUE, RESPECTIVELY.

Method	Input Resolution	LEVIR-CD			WHU-CD			CDD-CD		
		F1	OA	IoU	F1	OA	IoU	F1	OA	IoU
FC-EF [8]	256 x 256	83.40	98.39	71.53	69.37	97.61	53.11	66.93	93.28	50.30
FC-Siam-Diff [8]	256 x 256	86.31	98.67	75.92	58.81	95.63	41.66	70.61	94.95	54.57
FC-Siam-Conc [8]	256 x 256	83.69	98.49	71.96	66.63	97.04	49.95	75.11	94.95	60.14
DASNet [6]	256 x 256	79.91	94.32	66.54	70.50	97.29	54.41	92.70	98.20	86.39
DTCDSCN [47]	256 x 256	87.67	98.77	78.05	71.95	97.42	56.19	92.09	98.16	85.34
IFNet [7]	256 x 256	88.13	98.87	78.77	83.40	98.83	71.52	84.00	96.03	71.91
STANet [9]	256 x 256	87.30	98.66	77.40	82.32	98.52	69.95	84.12	96.13	72.22
MSTDSNet [12]	256 x 256	88.10	98.56	78.73	-	-	-	-	-	-
H-TransCD [14]	256 x 256	90.60	99.00	81.92	-	-	-	-	-	-
SNUNet [26]	256 x 256	88.16	98.82	78.83	83.50	98.71	71.67	83.40	96.23	72.11
BIT [16]	256 x 256	89.31	98.92	80.68	83.98	98.75	72.39	88.90	97.47	80.01
TransUNetCD [11]	256 x 256	91.11	-	83.67	93.59	-	84.42	97.17	-	94.50
ChangeFormer [10]	256 x 256	90.40	99.04	82.48	84.93	98.82	73.80	89.83	97.68	81.53
GeSANet [37]	256 x 256	90.05	99.01	81.90	63.02	96.20	46.00	95.14	98.83	90.73
ScratchFormer (ours)	256 x 256	91.68	99.16	84.63	91.87	99.37	84.97	97.88	99.50	95.85

TABLE II
COMPARISON OF PARAMETERS, INFERENCE TIME FOR SINGLE IMAGE PAIR, AND TRAIN TIME PER EPOCH WITH METHODS UTILIZING TRANSFORMER-BASED BACKBONE ON LEVIR-CD.

Method	Parameters (M)	Input Resolution	Train Time / Epoch (minutes)	Inference Time (ms)	IoU
Baseline	49.08	256 x 256	8.2	284	82.53
ChangeFormer [10]	41.03	256 x 256	8.1	249	82.48
ScratchFormer (ours)	36.95	256 x 256	7.9	268	84.63

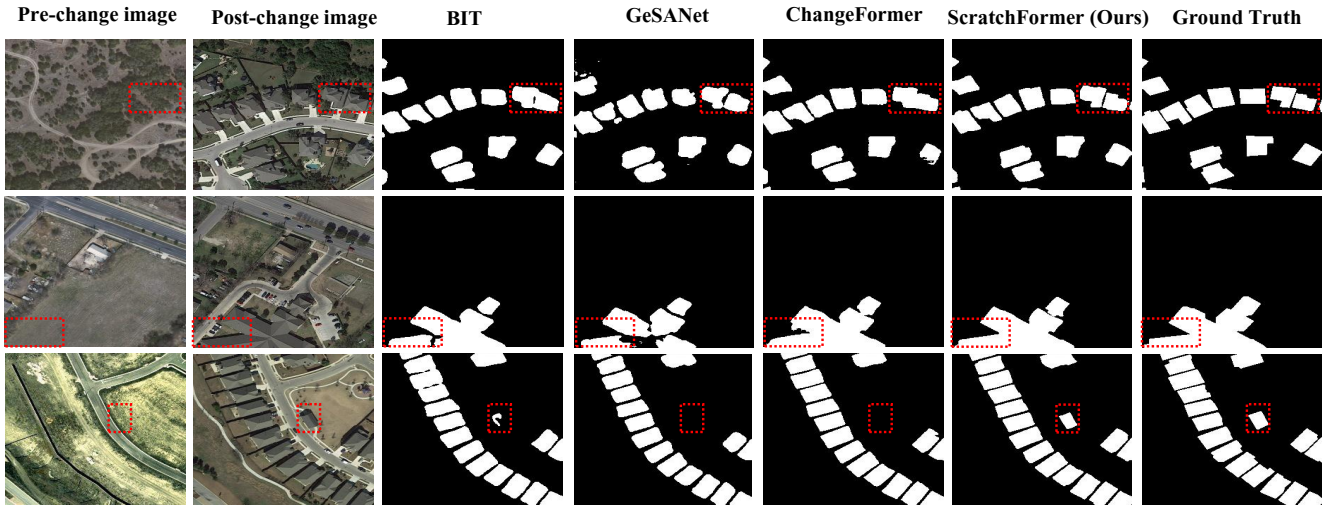

Fig. 6. Qualitative comparison on LEVIR-CD. We compare our ScratchFormer with BIT, GeSANet, and ChangeFormer. Our ScratchFormer provides improved CD performance by accurately detecting the correct changes (marked in red box) with clear boundaries, compared to existing methods.

methods in literature ChangeFormer [10] and TransUNetCD [11].

Comparison on WHU-CD: Here, we present the state-of-the-art comparison on the WHU-CD dataset (Tab. I). Among existing transformer-based methods, BIT [16] and TransUNetCD [11] achieve IoU scores of 72.39% and 84.42%, respectively. In comparison, our ScratchFormer which is trained from scratch through random initialization on this dataset achieves favorable performance against existing methods with an IoU score of 84.97%.

Comparison on CDD-CD: We also report results (Tab. I) on the CDD-CD dataset. Among CNN-based approaches, the DASNet [6] achieves IoU score of 86.39%. Among transformer-based CD methods, TransUNetCD [11] achieves an IoU score of 94.50%, which achieves this performance by employing an improved ResNet50 backbone. In comparison,

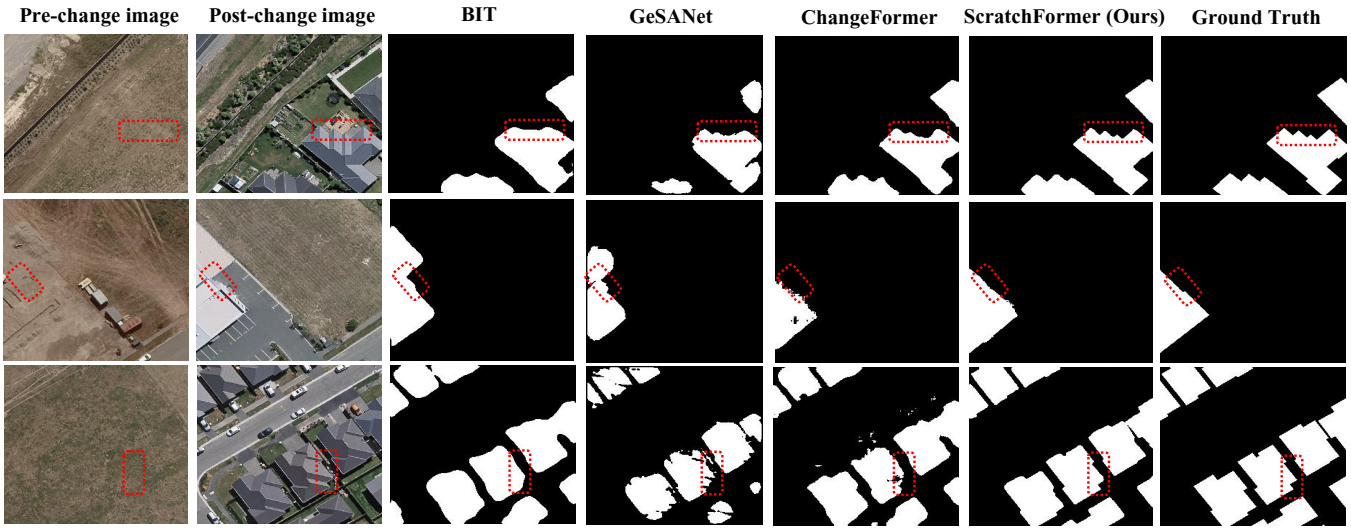

Fig. 7. Qualitative comparison on WHU-CD. We compare our ScratchFormer with BIT, GeSANet, and ChangeFormer. It is notable that our ScratchFormer provides improved CD performance by accurately detecting semantic changes with clear boundaries highlighted in red boxes.

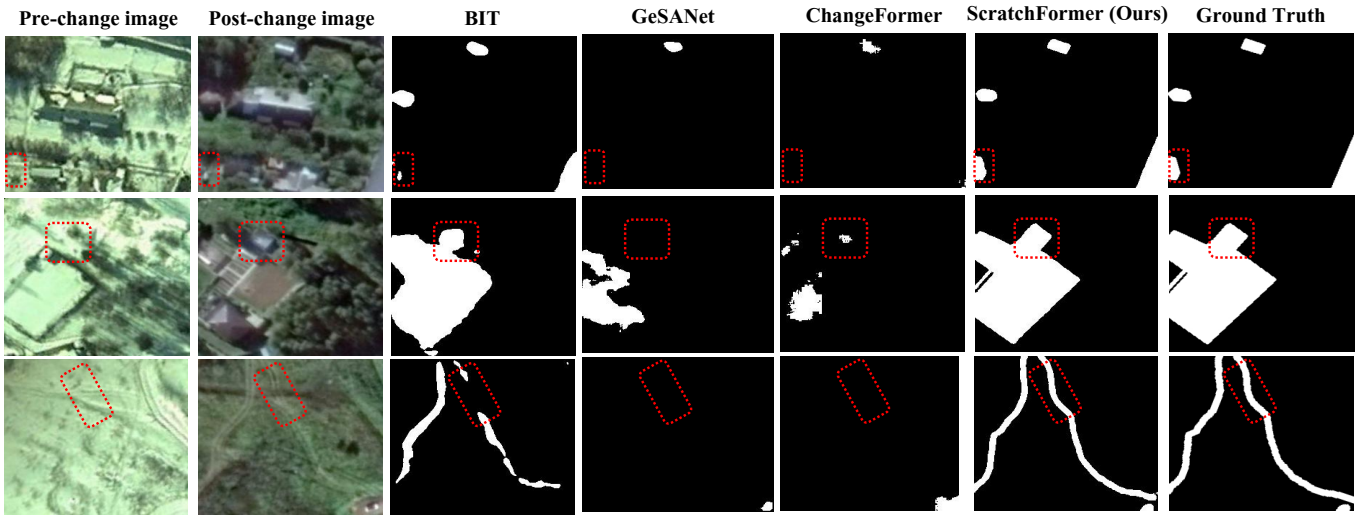

Fig. 8. Qualitative comparison on CDD-CD. We observe that our ScratchFormer better detects the semantic changes with clear boundaries between the pre- and post-change images.

TABLE III
STATE-OF-THE-ART COMPARISON ON DSIFN-CD DATASET IN TERMS OF F1, IOU, AND OA METRICS. FOR A FAIR COMPARISON, WE REPORT THE RESULTS BASED ON THE PUBLICLY AVAILABLE CODES OF STATE-OF-THE-ART METHODS. SCRATCHFORMER PERFORMS FAVORABLY AGAINST EXISTING METHODS AND ACHIEVES STATE-OF-THE-ART PERFORMANCE. THE BEST TWO RESULTS ARE IN RED AND BLUE, RESPECTIVELY.

Method	Input Resolution	DSIFN-CD		
		F1	OA	IoU
FC-Siam-Diff [8]	256 x 256	65.26	89.06	48.44
DTCDSCN [47]	256 x 256	65.29	88.14	48.46
BIT [16]	256 x 256	67.74	89.72	51.22
ChangeFormer [10]	256 x 256	69.50	90.56	53.26
GeSANet [37]	256 x 256	39.66	89.22	24.73
ScratchFormer (ours)	256 x 256	73.22	92.36	57.76

our ScratchFormer trained from scratch achieves an IoU score of 95.85%.

Comparison on DSIFN-CD: We compare our approach with both CNN-based and transformer-based state-of-the-art methods over DSIFN-CD. Tab. III presents the results. We observe that recent transformer-based methods achieve better F1 score. For instance, BIT [16] and ChangeFormer [10] achieve F1 scores of 67.74% and 69.50%, respectively. Our ScratchFormer outperforms these recent methods and achieves F1 score of 73.22%. Notably, ScratchFormer achieves absolute gains of 3.72% and 4.5% in terms of F1 and IoU compared to ChangeFormer [10]. It is worth mentioning that our approach here is trained from scratch without using any pre-training on another CD dataset. On this dataset, ScratchFormer sets a new state-of-the-art performance with a significant gain obtained in the challenging metrics.

Comparison on OSCD: Lastly, we present the results on OSCD dataset in Tab. V. We reproduce the numbers for FC-

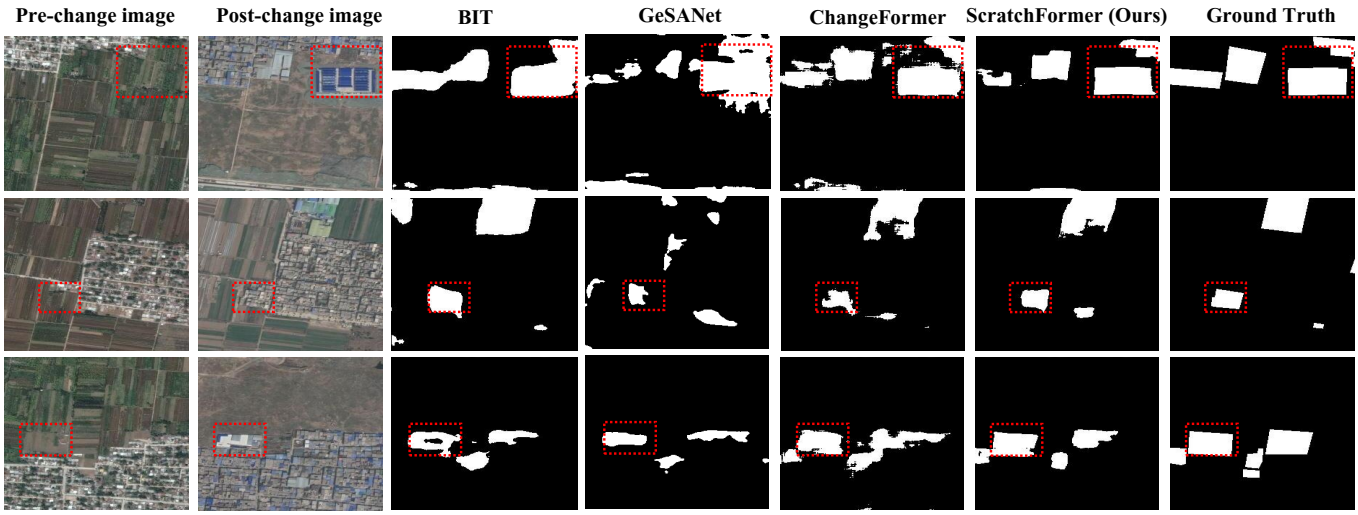

Fig. 9. Qualitative comparison on DSFIN-CD. We compare our ScratchFormer with BIT, GeSANet, and ChangeFormer. We observe our ScratchFormer to better detect the semantic changes (marked in red box) with clear boundaries between the pre- and post-change images, compared to other methods.

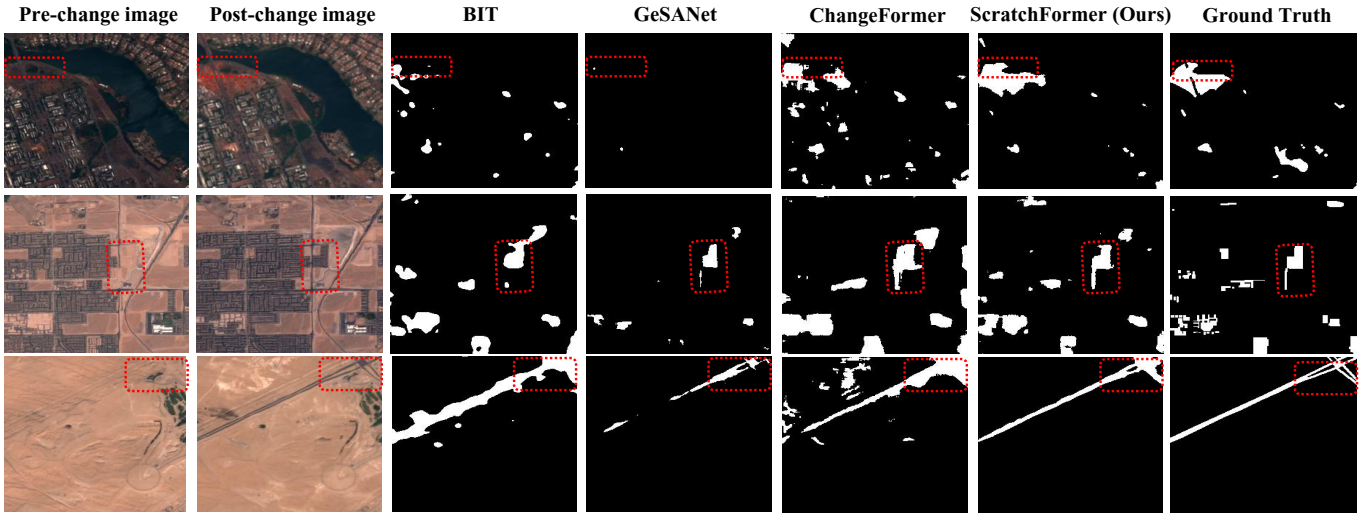

Fig. 10. Qualitative comparison on OSCD. We observe that our ScratchFormer better detects the semantic changes with clear boundaries between the pre- and post-change images highlighted in red boxes.

Siam-Diff [8], DTCDSCN [47], BIT [16], and ChangeFormer [10] Among the recent state-of-the-art methods, FC-Siam-Diff [8] achieves the best F1-score of 56.01%. However, Scratchformer being trained from scratch achieves a significantly better F1-score of 57.37% compared to the FC-Siam-Diff and sets new state-of-the-art results.

Comparison of Parameters and Time: We present the comparison of trainable parameters, the inference time for single image pair, and the time required to train the model for a single epoch of ScratchFormer with the other methods utilizing transformer based backbone. Tab. II shows that baseline has more parameters and inference time, and its performance in terms of IoU is inferior compared to the ScratchFormer. Besides, ChangeFormer [10] has slightly lower inference time while its trainable parameters and train time is higher compared to our method. Although our method has a slightly longer inference time, it has promising results in terms of all metrics, thereby providing a better trade-off with respect to performance and efficiency.

Qualitative Comparison: We present the qualitative comparison of our ScratchFormer with BIT [16], GeSANet [37], and ChangeFormer [10] in Figures 6, 7 from LEVIR-CD [9] and WHU-CD [23] examples, respectively. Figure 8 shows the qualitative comparison of our ScratchFormer with the transformer based methods [16], [10], [37] from CDD-CD [24] examples. Moreover, Figures 9 and 10 show the qualitative comparison of our ScratchFormer with BIT [16], GeSANet [37], and ChangeFormer [10] from DSIFN-CD [7], and OSCD [25] datasets, respectively. The results show that the proposed ScratchFormer is able to detect semantic changes occurring at multiple scales in complex scenes, enabling optimal CD performance when being trained from scratch directly on the target CD dataset. Furthermore, these qualitative analysis demonstrate the efficacy of our proposed ScratchFormer uti-

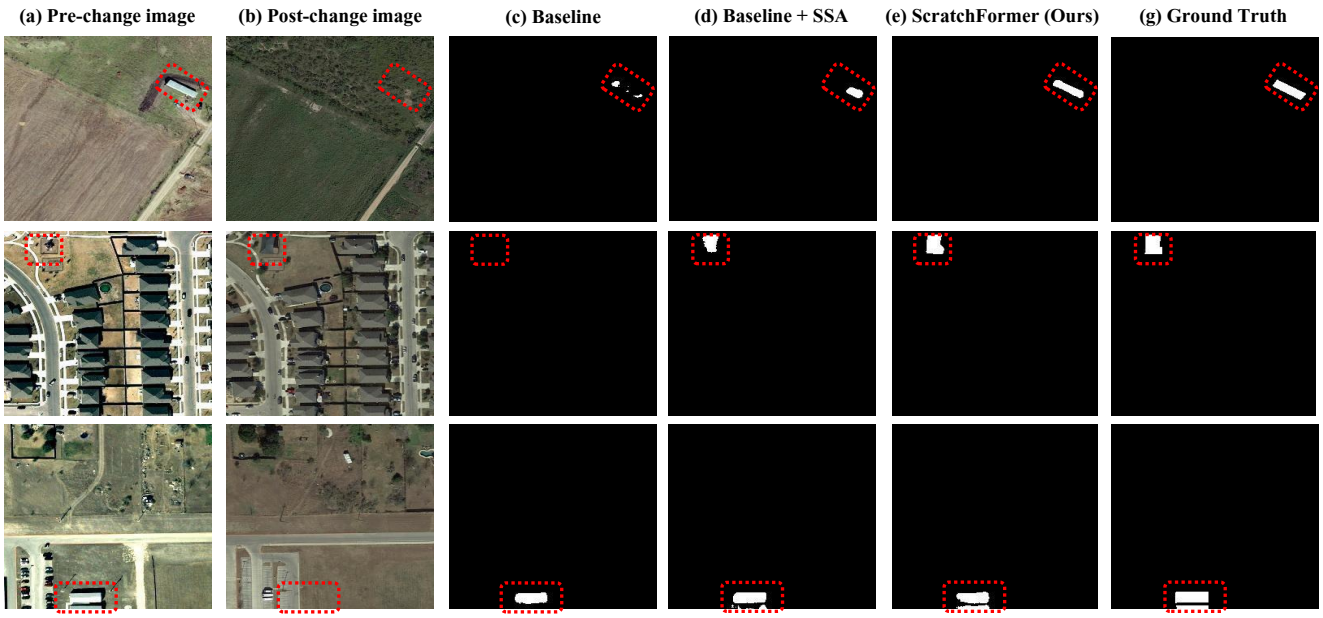

Fig. 11. Qualitative ablation study on LEVIR-CD. We compare our final ScratchFormer (e) which includes both contributions (SSA and CEFF). The change detection performance of baseline and SSA introduced to baseline is shown in (c) and (d), respectively. Our ScratchFormer (e) provides improved CD performance by accurately detecting the correct changes (marked in red box) with clear boundaries, demonstrating the effectiveness of our contributions.

TABLE IV
ABLATION STUDY ON THE LEVIR-CD DATASET. HERE, WE SHOW THE IMPACT OF INTEGRATING OUR CONTRIBUTIONS TO THE BASELINE. † DENOTES THAT THE MODEL IS PRE-TRAINED FIRST ON ANOTHER CD DATASET AND THEN FINETUNED ON THE TARGET CD DATASET. THE INTEGRATION OF OUR SSA (ROW 5) INTO THE BASELINE (ROW 4) LEADS TO CONSISTENT GAIN IN PERFORMANCE. OUR FINAL APPROACH SCRATCHFORMER (ROW 6) WHICH COMPRISES BOTH SSA AND CEFF ACHIEVES A SIGNIFICANT IMPROVEMENT IN PERFORMANCE OVER THE BASELINE. HERE, WE ALSO REPORT CHANGEFORMER WITH AND WITHOUT PRE-TRAINING. THE BEST TWO RESULTS ARE IN RED AND BLUE, RESPECTIVELY.

Method	LEVIR-CD		
	F1	OA	IoU
ChangeFormer [10] †	90.40	99.04	82.48
ChangeFormer [10]	84.97	98.52	73.86
Baseline †	90.65	99.06	82.89
Baseline	90.43	99.05	82.53
Baseline + SSA (Sec. IV-C)	91.08	99.09	83.62
Baseline + SSA+ CEFF (**ScratchFormer**)	91.68	99.16	84.63

TABLE V
FOR A FAIR COMPARISON, WE PROVIDE A STATE-OF-THE-ART COMPARISON OF OSCD DATASET. WE REPORT THE RESULTS IN TERMS OF F1, IOU, AND OA METRICS. SCRATCHFORMER PERFORMS SIGNIFICANTLY BETTER AGAINST EXISTING METHODS AND ACHIEVES STATE-OF-THE-ART PERFORMANCE. THE BEST TWO RESULTS ARE IN RED AND BLUE, RESPECTIVELY.

Method	Input Resolution	OSCD		
		F1	OA	IoU
FC-Siam-Diff [8]	256 x 256	56.01	96.69	38.90
DTCDSCN [47]	256 x 256	43.57	97.13	27.85
BIT [16]	256 x 256	48.97	96.50	32.42
ChangeFormer [10]	256 x 256	49.23	94.93	32.65
GeSANet [37]	256 x 256	35.99	97.12	21.94
ScratchFormer (ours)	256 x 256	57.37	97.33	40.22

lizing novel shuffled sparse attention which focuses on sparse informative regions to capture the inherent characteristics of the CD data.

C. Ablation Study

Here, we present ablation study to validate the effectiveness of our contributions over LEVIR-CD dataset. Tab. IV shows the baseline comparison. The baseline approach (Sec. III-A) when trained from scratch using random initialization achieves IoU score of 82.53% (row 4) over LEVIR-CD dataset. The results of the baseline approach are improved to 82.89% (row 3) when first pre-training it on DSIFN-CD and then finetuning it on the LEVIR-CD (target) dataset. When integrating our SSA layer (Sec. IV-C) into the baseline, the results are improved to 83.62% in terms of IoU score (row 5). Our final ScratchFormer which includes both contributions (SSA and CEFF) and trained from scratch leads to a significant improvement in performance by achieving an IoU score of 84.63%. These results demonstrate the effectiveness of our contributions. In addition to the baseline comparison, we also report the results of ChangeFormer using both pre-training and training from scratch. Our ScratchFormer achieves consistent gain in performance on all three metrics over the ChangeFormer.

We further perform an experiment to compare our CEFF module with standard addition, subtraction, and concatenation based techniques. Here, addition, subtraction, and concatenation are performed for $\hat{F}^i_{pre}$ and $\hat{F}^i_{post}$, and passed to two convolutional layers. Tab. VII shows the comparison. Our CEFF that utilizes feature channel re-weighting achieves superior performance compared to these techniques.

Shuffled Sparse Features: The calculation method to predict

TABLE VI
COMPARISON OF THE SPARSITY γ OVER LEVIR-CD DATASET. THE SPARSITY $\gamma = 4$ ACHIEVES SUPERIOR PERFORMANCE. THE BEST RESULTS ARE IN BOLD.

γ	LEVIR-CD		
	F1	OA	IoU
γ=2	91.49	99.14	84.32
γ=4	**91.68**	**99.16**	**84.63**
γ=8	91.56	99.15	84.44

TABLE VII
COMPARISON OF CEFF WITH THE SUBTRACTION, ADDITION, AND CONCATENATION-BASED TECHNIQUES ON LEVIR-CD. CEFF ACHIEVES SUPERIOR PERFORMANCE ON ALL METRICS. THE BEST TWO RESULTS ARE IN RED AND BLUE, RESPECTIVELY.

Method	LEVIR-CD		
	F1	OA	IoU
Difference module with Subtraction	90.74	99.07	83.05
Difference module with Addition	91.02	99.10	83.52
Difference module with Concatenation	91.08	99.09	83.62
CEFF	91.68	99.16	84.63

the offsets is adapted from deformable convolutional network [45]. Our approach then employs sparse sub-sampling instead of a dense sub-sampling. We empirically observe our approach to achieve superior performance, compared to using a dense sub-sampling with shuffled locations using the computed offsets (Our approach: 84.63% vs. dense sub-sampling: 83.37% on LEVIR-CD in terms of IoU score). We further conjecture this improvement to be likely due to effectively learning a rich feature representations by attending to sparse informative regions in remote sensing CD images. In contrast, the dense sub-sampling on uniformly sampled dense patches is likely to have difficulties to learn a rich feature representation encoding diverse shape objects with inconsistent appearance in remote sensing scenes having sparse informative regions.

Sparsity Factor: We also conduct an experiment to estimate the optimal sparsity of our SSA by varying the sparsity factor γ (2, 4, and 8) as shown in Table VI. We observe setting the value of γ to 4 to provide optimal performance for LEVIR-CD dataset. Therefore, we fix the γ and use the same value throughout our experiments.

VI. CONCLUSION

We propose a transformers-based Siamese architecture, named ScratchFormer, for the problem of remote sensing change detection. Our ScratchFormer introduces shuffled sparse attention to effectively capture the inherent characteristics when training from scratch. We further introduce a change-enhanced feature fusion module to perform per-channel feature weighting to enhance the relevant semantic changes, while suppressing the noisy ones. We validate our approach by conducting extensive experiments on multiple commonly used change detection benchmarks with different set of challenges. For instance, LEVIR-CD and WHU-CD datasets present different challenges such as building shadows, color variations, vegetation changes, and various types of buildings having irregular shapes and sizes, whereas the CDD-CD dataset poses challenges in terms of accurate boundary delineation likely due to different factors including, image resolution, sensor limitations, and the nature of the changes due to season-varying image acquisition. Our approach performs favorably against existing change detection methods on all these datasets. A potential future research direction is to further explore accurate boundary delineation particularly in scenarios with season-varying images along with the generalizability of the transformers-based remote sensing change detection at provincial-scale [48], [49]. Another future direction is to investigate the problem of change detection in natural images and medical imaging.

ACKNOWLEDGEMENTS

The authors would like to thank and express their deepest gratitude to Mohamed bin Zayed University of Artificial Intelligence for the continuous support throughout the research journey.

Mubashir Noman is a PhD student in the computer vision department at Mohamed bin Zayed University of Artificial Intelligence (MBZUAI) located in Abu Dhabi, UAE.

His research interests include image processing, computer vision, and remote sensing.

Mustansar Fiaz received the bachelor's degree from the Pakistan Institute of Engineering and Applied Science (PIEAS), Islamabad, Pakistan, in 2011, the master's degree from Sejong University, Seoul, South Korea, in 2016, and the Ph.D. degree from Kyungpook National University, Daegu, South Korea, in 2021.

He is a Researcher with IBM Research, UAE. Before joining IBM Research, he was a Research Associate with the Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE. Before joining MBZUAI, he was a Senior AI Software Engineer in the AI Industry. His research interests include computer vision, remote sensing, medical imaging, and vision and language.

Dr. Fiaz for his Ph.D. thesis received an Outstanding Research CSE Thesis Award.

Hisham Cholakkal received the Ph.D. degree in computer vision from Nanyang Technological University, Singapore, in 2016.

He is an Assistant Professor with the MBZ University of Artificial Intelligence, Abu Dhabi, UAE. Prior to joining MBZUAI, he worked as a Research Scientist with the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi. Before joining IIAI, he was a Senior Technical Lead in the Computer Vision and Deep Learning Research Team with Mercedes-Benz Research and Development, Bengaluru, India. He has also worked as a Researcher with BEL-Central Research Laboratory, Bengaluru, and Advanced Digital Sciences Center, Singapore. He has diverse experiences across fundamental research, teaching, and product development in industry. He has several years of experience in leading research teams involved in commercial and fundamental research. In addition to authoring top-tier research articles and patents, he developed several computer vision frameworks that are successfully released as commercial products in various industries. His recent research interests include object detection, image and video segmentation, object counting, image classification, pedestrian detection, person search, human–pose estimation, human–object interaction detection, activity recognition, crowd counting, few-shot/zero-shot learning, weakly supervised learning, and AI for style imitation and AI for creativity.

Dr. Cholakkal has served as a Program Committee Member for several top conferences, including CVPR, ICCV, NeurIPS, ICLR, and ECCV.

Sanath Narayan received the PhD degree from the Indian Institute of Science in 2016. He is a research scientist with the Technology Innovation Institute, Abu Dhabi. He previously worked as a research scientist with Inception Institute of Artificial Intelligence and as a senior technical lead with Mercedes-Benz R&D India. He has served as a program committee member for several premier conferences including CVPR, ICCV and ECCV. He has been recognized as an outstanding/top reviewer multiple times with these conferences. His thesis was awarded the best doctoral symposium paper award at ICVGIP 2014. His research interests include computer vision and machine learning

Salman Khan received the Ph.D. degree from the University of Western Australia, Perth, WA, Australia, in 2016.

He is an Associate Professor with the MBZ University of Artificial Intelligence, Abu Dhabi, UAE. He has been an Adjunct Faculty Member with Australian National University, Canberra, ACT, Australia, since 2016. His research interests include computer vision and machine learning.

Dr. Khan served as a Program Committee Member for several premier conferences, including CVPR, ICCV, ICLR, ECCV, and NeurIPS. He has been awarded the Outstanding Reviewer Award at CVPR Multiple Times, won the Best Paper Award at Ninth ICPRAM 2020, and Second Prize in the NTIRE Image Enhancement Competition at CVPR 2019. His thesis received an honorable mention on the Dean's List Award.

Fahad Shahbaz Khan (Senior Member, IEEE) received the M.Sc. degree in intelligent systems design from the Chalmers University of Technology, Gothenburg, Sweden, in 2007, and the Ph.D. degree in computer vision from the Autonomous University of Barcelona, Barcelona, Spain, in 2011.

He is currently a Full Professor and the Deputy Department Chair of Computer Vision with MBZUAI, Abu Dhabi, UAE. He also holds a Faculty Position (Universitetslektor + Docent) with Computer Vision Laboratory, Linköping University, Linköping, Sweden. From 2018 to 2020, he worked as a Lead Scientist with the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi. His research interests include a wide range of topics within computer vision and machine learning, such as object recognition, object detection, action recognition, and visual tracking. He has published articles in high-impact computer vision journals and conferences in these areas.

Dr. Khan serves as a Regular Program Committee Member for leading computer vision conferences, such as CVPR, ICCV, and ECCV. He has achieved top ranks on various international challenges (Visual Object Tracking (VOT): 1st 2014 and 2018, 2nd 2015, 1st 2016; VOT-TIR: 1st 2015 and 2016; OpenCV Tracking: 1st 2015; 1st PASCAL VOC 2010).

REFERENCES

[1] M. Kucharczyk and C. H. Hugenholtz, "Remote sensing of natural hazard-related disasters with small drones: Global trends, biases, and research opportunities," *Remote Sensing of Environment*, vol. 264, p. 112577, 2021. 1

[2] J. Yin, J. Dong, N. A. Hamm, Z. Li, J. Wang, H. Xing, and P. Fu, "Integrating remote sensing and geospatial big data for urban land use mapping: A review," *International Journal of Applied Earth Observation and Geoinformation*, vol. 103, p. 102514, 2021. 1

[3] L. M. Fonseca, T. S. Körting, H. d. N. Bendini, C. D. Girolamo-Neto, A. K. Neves, A. R. Soares, E. C. Taquary, and R. V. Maretto, "Pattern recognition and remote sensing techniques applied to land use and land cover mapping in the brazilian savannah," *Pattern recognition letters*, vol. 148, pp. 54–60, 2021. 1

[4] D. Wen, X. Huang, F. Bovolo, J. Li, X. Ke, A. Zhang, and J. A. Benediktsson, "Change detection from very-high-spatial-resolution optical remote sensing images: Methods, applications, and future directions," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 4, pp. 68–101, 2021. 1

[5] M. Zhang, G. Xu, K. Chen, M. Yan, and X. Sun, "Triplet-based semantic relation learning for aerial remote sensing image change detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 2, pp. 266–270, 2018. 1

[6] J. Chen, Z. Yuan, J. Peng, L. Chen, H. Huang, J. Zhu, Y. Liu, and H. Li, "Dasnet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 1194–1206, 2020. 1, 3, 8

[7] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, and G. Liu, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 183–200, 2020. 1, 2, 3, 7, 8, 10

[8] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 4063–4067. 1, 8, 9, 10, 11

[9] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sensing*, vol. 12, no. 10, p. 1662, 2020. 1, 2, 3, 7, 8, 10

[10] W. G. C. Bandara and V. M. Patel, "A transformer-based siamese network for change detection," *arXiv preprint arXiv:2201.01293*, 2022. 1, 2, 3, 4, 6, 7, 8, 9, 10, 11

[11] Q. Li, R. Zhong, X. Du, and Y. Du, "Transunetcd: A hybrid transformer network for change detection in optical remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–19, 2022. 1, 2, 3, 7, 8

[12] F. Song, S. Zhang, T. Lei, Y. Song, and Z. Peng, "Mstdsnet-cd: Multiscale swin transformer and deeply supervised network for change detection of the fast-growing urban regions," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022. 1, 2, 3, 8

[13] G. Wang, B. Li, T. Zhang, and S. Zhang, "A network combining a transformer and a convolutional neural network for remote sensing image change detection," *Remote Sensing*, vol. 14, no. 9, p. 2228, 2022. 1

[14] Q. Ke and P. Zhang, "Hybrid-transcd: A hybrid transformer remote sensing image change detection network via token aggregation," *ISPRS International Journal of Geo-Information*, vol. 11, no. 4, p. 263, 2022. 1, 2, 3, 7, 8

[15] M. Noman, M. Fiaz, H. Cholakkal, S. Khan, and F. S. Khan, "Elgc-net: Efficient local–global context aggregation for remote sensing change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–11, 2024. 1

[16] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021. 1, 2, 3, 6, 7, 8, 9, 10, 11

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. 2, 3, 4, 6

[18] W. Shi, M. Zhang, R. Zhang, S. Chen, and Z. Zhan, "Change detection based on artificial intelligence: State-of-the-art and challenges," *Remote Sensing*, vol. 12, no. 10, p. 1688, 2020. 2, 3

[19] J. Wang, Y. Zhong, and L. Zhang, "Change detection based on supervised contrastive learning for high-resolution remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023. 2

[20] Y. Zhang, Y. Zhao, Y. Dong, and B. Du, "Self-supervised pre-training via multi-modality images with transformer for change detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 2

[21] Y. Zhang, W. Li, Y. Wang, Z. Wang, and H. Li, "Beyond classifiers: Remote sensing change detection with metric learning," *Remote Sensing*, vol. 14, no. 18, p. 4478, 2022. 2

[22] E. Guo, X. Fu, J. Zhu, M. Deng, Y. Liu, Q. Zhu, and H. Li, "Learning to measure change: Fully convolutional siamese metric networks for scene change detection," *arXiv preprint arXiv:1810.09111*, 2018. 2

[23] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 574–586, 2019. 2, 7, 10

[24] M. Lebedev, Y. V. Vizilter, O. Vygolov, V. A. Knyaz, and A. Y. Rubis, "Change detection in remote sensing images using conditional adversarial networks," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2018. 2, 7, 10

[25] R. C. Daudt, B. L. Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral earth observation using convolutional neural networks," 2018. 2, 7, 10

[26] S. Fang, K. Li, J. Shao, and Z. Li, "Snunet-cd: A densely connected siamese network for change detection of vhr images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021. 3, 8

[27] H. Jiang, X. Hu, K. Li, J. Zhang, J. Gong, and M. Zhang, "Pga-siamnet: Pyramid feature-based attention-guided siamese network for remote sensing orthoimagery building change detection," *Remote Sensing*, vol. 12, no. 3, p. 484, 2020. 3

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015. 3

[29] 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 Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2021. 3

[30] B. Hou, Y. Wang, and Q. Liu, "Change detection based on deep features and low rank," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 12, pp. 2418–2422, 2017. 3

[31] J. Xu, C. Luo, X. Chen, S. Wei, and Y. Luo, "Remote sensing change detection based on multidirectional adaptive feature fusion and perceptual similarity," *Remote Sensing*, vol. 13, no. 15, p. 3053, 2021. 3

[32] W. Wang, X. Tan, P. Zhang, and X. Wang, "A cbam based multiscale transformer fusion approach for remote sensing image change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 6817–6825, 2022. 3

[33] Y. Liang, C. Zhang, and M. Han, "Rasrnet: An end-to-end relation-aware semantic reasoning network for change detection in optical remote sensing images," *IEEE Transactions on Instrumentation and Measurement*, 2023. 3

[34] A. A. Aleissaee, A. Kumar, R. M. Anwer, S. Khan, H. Cholakkal, G.-S. Xia, and F. S. Khan, "Transformers in remote sensing: A survey," *Remote Sensing*, vol. 15, no. 7, 2023. 3

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. 3

[36] Y. Zhou, F. Wang, J. Zhao, R. Yao, S. Chen, and H. Ma, "Spatial-temporal based multihead self-attention for remote sensing image change detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6615–6626, 2022. 3

[37] X. Zhao, K. Zhao, S. Li, and X. Wang, "Gesanet: Geospatial-awareness network for vhr remote sensing image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 3, 8, 9, 10, 11

[38] L. Hu, J. Liu, and L. Xiao, "A total variation regularized bipartite network for unsupervised change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2022. 3

[39] S. Fang, K. Li, and Z. Li, "Changer: Feature interaction is what you need for change detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 3

[40] C. Zhang, L. Wang, S. Cheng, and Y. Li, "Swinsunet: Pure transformer network for remote sensing image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022. 3

[41] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022. 3

[42] D. Wang, X. Chen, N. Guo, H. Yi, and Y. Li, "Stcd: efficient siamese transformers-based change detection method for remote sensing images," *Geo-spatial Information Science*, pp. 1–20, 2023. 3

[43] Y. Teng, S. Liu, W. Sun, H. Yang, B. Wang, and J. Jia, "A vhr bi-temporal remote-sensing image change detection network based on swin transformer," *Remote Sensing*, vol. 15, no. 10, p. 2645, 2023. 3

[44] D. Hong, C. Qiu, A. Yu, Y. Quan, B. Liu, and X. Chen, "Multi-task learning for building extraction and change detection from remote sensing images," *Applied Sciences*, vol. 13, no. 2, p. 1037, 2023. 3

[45] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773. 6, 12

[46] T. Yan, Z. Wan, and P. Zhang, "Fully transformer network for change detection of remote sensing images," *arXiv preprint arXiv:2210.00757*, 2022. 6

[47] 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 Geoscience and Remote Sensing Letters*, vol. 18, no. 5, pp. 811–815, 2020. 8, 9, 10, 11

[48] X. L. Y. Z. G. X. Da He, Qian Shi and L. Zhang, "Generating annual high resolution land cover products for 28 metropolises in china based on a deep super-resolution mapping network using landsat imagery," *GIScience & Remote Sensing*, vol. 59, no. 1, pp. 2036–2067, 2022. [Online]. Available: https://doi.org/10.1080/15481603.2022.2142727 12

[49] Q. Shi, M. Liu, A. Marinoni, and X. Liu, "Ugs-1m: fine-grained urban green space mapping of 31 major cities in china based on the deep learning framework," *Earth System Science Data*, vol. 15, no. 2, pp. 555–577, 2023. [Online]. Available: https://essd.copernicus.org/articles/15/555/2023/ 12