Change Detection Enhanced by Spatial-Temporal Association for Bare Soil Land Using Remote Sensing Images

Sasha Wu, Yalan Liu, Shufu Liu, Dacheng Wang, Linjun Yu, and Yuhuan Ren

***Abstract*—As dust source bare soil land (BSL) contributes to air pollution and affects the photosynthesis of green plants and carbon absorption, it is the objective of this study to develop an approach for monitoring the changes of BSL using remote sensing technology. Unlike other land use/cover types, the classification of BSL as well as its change detection is often ignored. For traditional convolutional neural networks, deep layers cause a long range between input and output, inevitably leading to the loss of information and computational costs. To alleviate this problem, transformer is available to model the global dependencies. Bitemporal association, which is described as subtraction or attention mechanism, is not fully considered by current methods. Therefore, we proposed a spatial-temporal association enhanced mobile-friendly vision transformer (STAE-MobileVIT) for change detection of high-resolution images with light weight and high efficiency. On the one hand, a temporal association enhanced MobileVIT block is employed to strengthen the association of bitemporal images during feature extraction. On the other hand, a multiscale feature difference aggregator enhanced by spatial association is designed to fuse semantic and detailed information. Since the lack of binary change detection dataset for BSL, we established a small dataset named BSL-CD, consisting of 1083 pairs of 0.8 m bitemporal images with the size of 256 × 256 pixels, along with the corresponding labels. The experiments on BSL-CD show that our light-weight model surpass seven common methods by 3.48, 5.05, and 1.44 percent on F1, IoU, and OA, which proves the efficiency and accuracy of STAE-MobileVIT.**

***Index Terms*—Bare soil land (BSL), change detection, mobileVIT, remote sensing images, spatial-temporal association.**

I. Introduction

DUST source bare soil land (BSL) increases due to construction, demolition, and residential relocation for urbanization, as well as secondary plowing and reclamation. BSL often contributes to air pollution, which does harm to human health [1], and impacts the photosynthesis of green plants and carbon absorption, water loss, and soil erosion [2]. It is essential to obtain dynamic changes accurately and efficiently in order to support environmental governance.

"BSL" is defined as soil-covered land on surface, primarily devoid of vegetation cover according to Chinese National Standard for Land Use Classification [3]. However, BSL is mostly neglected in many classification systems for land use/cover, and is classified as "other land categories" together with bare rock, gravel land, and sandy land. The areas covered by exposed soil and rare vegetation are at high risk of dust pollution. To meet the requirement of environmental governance, we take this kind of "bare soil land" as BSL in this study. With the development in large and medium-sized cities, BSL has the potential to transform into buildings, concrete floor, forest, grassland, cropland, and water and dust-proof nets. Conversely, the reverse is also true. We consider these changes related to BSL as our target areas to be extracted. Among the abovementioned changes, areas of critical concern include the emerging BSL, dust-proof nets, the grassland, and cropland with seasonal variations. BSL serves as the main source of dust pollution and poses a great challenge to urban environmental governance in China. However, existing products fail to satisfy the requirements for practical application in accuracy and efficiency. For this reason, more and more attention is paid to BSL.

In terms of change detection for multispectral remote sensing images, the traditional methods mainly extract simple features to obtain pixel-level or object-oriented [4] change results, for instance, band operation [5], [6], [7], image transformation [8], [9], [10], and multisource information assisted methods [11]. However, it is difficult for them to distinguish BSL from impervious surface and buildings in images. With the advent of machine learning methods, support vector machine [12], random forest [13], and multilayer perception (MLP) [14] are applied to improve the precision of change detection. Nevertheless, it is tough for them to extract changes automatically from large areas and apply to other different regions, since they rely on manually generated and selected features from specific areas, which limits the contextual scope and requires prior knowledge. With the increasing availability of high-spatial resolution remote sensing images, which provide more and more information of land surface and fine-grained changes, the abovementioned methods encounter challenges in processing a large volume of

Manuscript received 11 June 2023; revised 4 September 2023; accepted 16 October 2023. Date of publication 23 October 2023; date of current version 23 November 2023. This work was supported by the Project of Dynamic Remote Sensing Monitoring of Bare Soil in Daxing District, Beijing, China under Grant E2H2110302. *(Corresponding authors: Yalan Liu; Yuhuan Ren.)*

Sasha Wu and Yalan Liu are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: wusasha21@mails.ucas.ac.cn; liuyl@aircas.ac.cn).

Shufu Liu, Dacheng Wang, Linjun Yu, and Yuhuan Ren are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China (e-mail: liusf@aircas.ac.cn; wangdc@aircas.ac.cn; yulj201831@aircas.ac.cn; renyh@aircas.ac.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3326958

images efficiently. With the development of artificial intelligence and computer vision (CV) techniques, deep learning models are popularly applied to change detection for remote sensing images. These models excel in extracting spectral, textual, geometrical boundary, and context-rich features, enabling automatic change detection. Most deep learning models for change detection are modified from classical networks, such as fully convolutional network (FCN) [15] and UNet [16]. Zhang et al. [17] applied multiscale supervision to improve the precision and Chen and Shi [18] employed the spatial-temporal attention mechanism to prioritize the bitemporal association with a relatively large model. The structures of the above change detection models can be basically divided into one-stream and two-stream models [19]. For the one-stream model, the corresponding bands are first selected from the multibands of bitemporal images. Subsequently, the selected channels are either connected [20] or subtracted [21]. Finally, the one-stream model is trained as a semantic segmentation model for the extraction of changed areas. For the two-stream model, Siamese networks are utilized to extract bitemporal features and detect areas of change. However, the above models based on convolutional neural network (CNN) cause the loss of information during multilayer encoding and decoding. With deeper layers, the number of channels increases, resulting in more parameters, a larger model, and longer training time. As one of the innovative deep learning methods designed for natural language processing, transformer [22] whose self-attention mechanism can capture global dependencies availably has been recently introduced to CV tasks like image classification [23] and semantic segmentation [24]. Hence, Chen et al. [25] and Bandara and Patel [26] applied transformer to encode the input bitemporal images into context-rich tokens for change detection.

Although significant progress has been made in change detection for remote sensing images, there are still some challenges that persist in BSL change detection. First, most of the models for change detection are based on classical networks with a large volume, but few models are specifically developed for small datasets. If a small dataset is trained using a large-scale model like Deeplabv3+ [27] with the Xception backbone, it is easy to cause the problem of overfitting. For datasets with a small sample size, light-weight networks like ShuffleNet [28], MobileNet [29], and mobile-friendly vision transformer (MobileVIT) [30] have fewer parameters and deliver impressive performance in image classification and semantic segmentation. However, they have not been widely applied to change detection tasks. Second, it is not sufficient for BSL to extract bitemporal features with the only connection of weight shared Siamese networks. Unlike other land use/cover types, the change detection results of BSL are prominently influenced by pseudochanges result from seasonal variations, atmospheric conditions, acquisition time, and images perspectives. This impact is particularly pronounced in areas covered by growing or withered grass. Furthermore, the spatial scales of BSL could be considerably various, different from the common objects for change detection, like buildings in similar scales. Thus, it is crucial to strengthen the information interaction between bitemporal images and explore the global relations among pixels in spatial-temporal domain. However, few methods are designed for the change characteristics of BSL. Third, BSL change detection faces significant limitations due to the scarcity of available datasets. While there are public datasets for change detection in specific domains, such as binary LEVIR-CD [18] for buildings and CLCD [31] for cropland, SECOND [32], and Hi-UCD [33] for semantic change detection, there is a notable lack of attention and dedicated datasets specifically focused on BSL change detection.

To address these issues, we proposed a spatial-temporal association enhanced mobile-friendly vision transformer (STAE-MobileVIT) and established a dataset (BSL-CD) for change detection of BSL using high-resolution images. The encoder of STAE-MobileVIT is Siamese networks. A light-weight CNN backbone (MobileNetV2) [34] is applied to extract multiscale features, and a temporal association enhanced MobileVIT (TAE-MobileVIT) block is used to enhance the association of bitemporal images. The attention mechanism is incorporated into a spatial association enhanced (SAE) decoder, where feature difference maps from multiple scales are fused to generate context-rich feature maps. Finally, a shallow CNN is applied to produce pixel-level predictions for changes. BSL-CD consists of 1083 pairs of bitemporal images annotated with different BSL changes, which can provide a benchmark for training deep learning models in studies on BSL change detection.

The contributions of our work can be summarized as follows.

1) An efficient and light-weight CNN-transformer model STAE-MobileVIT is proposed for BSL change detection. We apply light-weight backbones MobileNetV2 and MobileVIT to lighten the proposed model and to adapt to small datasets. In addition, a TAE-MobileVIT block is used to strengthen the temporal association of bitemporal images.
2) A high-resolution dataset BSL-CD is provided for future studies, which contains 1083 pairs of bitemporal images with 0.8 m spatial resolution and size of 256×256 pixels, along with the corresponding binary changed labels.
3) The effectiveness and efficiency of STAE-MobileVIT are validated through comparative experiments on BSL-CD dataset with seven common change detection models including six CNN-based models and a traditional transformer-based model. The proposed method exceeds the second-ranked method by 3.48, 5.05, and 1.44 percent on F1, IoU, and OA.

The rest of this article is organized as follows. Section II introduces the details of the proposed methodology. Section III presents BSL-CD and the experimental settings. Some experimental results are reported and analyzed in Section IV. The discussion is given in Section V. Finally, Section VI concludes this article.

II. Methodology

As shown in Fig. 1, the overall structure of STAE- MobileVIT contains three components: a feature extractor, a SAE multiscale feature difference aggregator and a prediction head. The details of each part are introduced in the following.

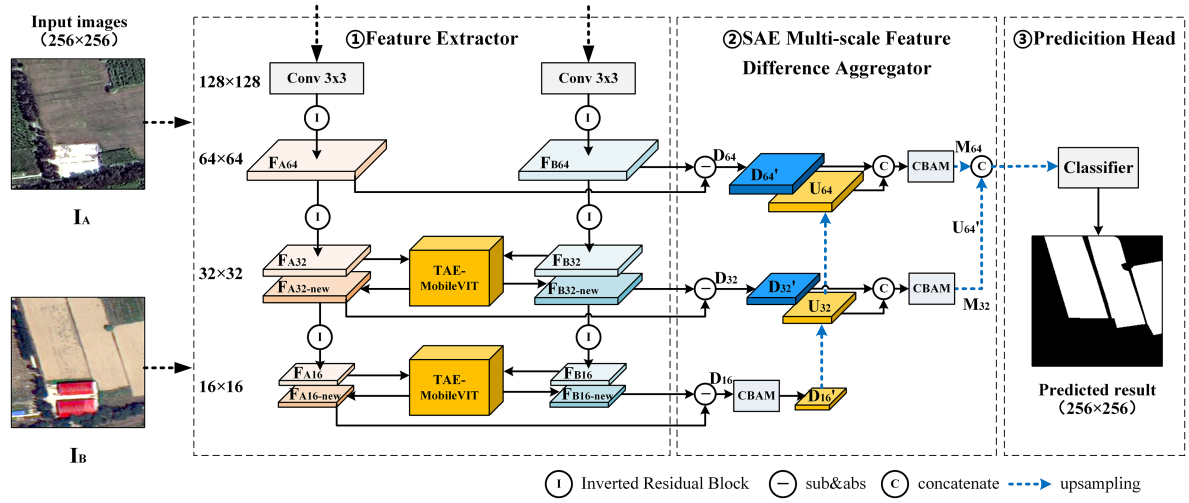

Fig. 1. Illustration of the proposed STAE-MobileVIT.

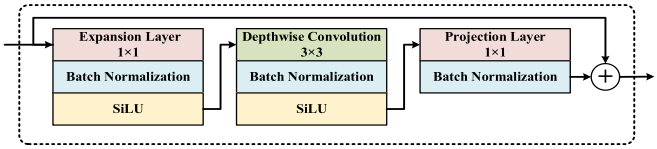

Fig. 2. Structure of the inverted residual block of MobileNetV2.

A. Feature Extractor

STAE-MobileVIT employs Siamese CNN-transformer networks modified from MobileVIT to extract multiscale feature maps of bitemporal images. First, we remove the inverted residual block of MobileNetV2 in the last layer based on the understanding that smaller feature maps contain less information. Parameters and the volume of the model also increase with deeper convolutional layers. Second, in order to enhance the temporal association, we concatenate bitemporal token sets in TAE-MobileVIT block instead of enhancing the features separately.

The procedure of the feature extractor is introduced as following. The input bitemporal RGB images are cropped into size of 256 × 256 pixels, which facilitates the balance of target and background pixels within smaller-sized samples. The first convolutional layer with a kernel size of 3 and a stride of 2 is used to extract half-size shallow features, followed a batch normalization [35] layer and a sigmoid-weighted linear unit (SiLU) function [36]. Then, we use three inverted residual blocks to extract feature maps with the size of 64 × 64, 32 × 32, and 16 × 16. After obtaining the features with the size of 32 × 32 and 16 × 16, a TAE-MobileVIT block is employed in each scale to strengthen the temporal association and enhance the bitemporal features. According to the parametric analysis in Section V, the depths of transformer are set to 2 and 4, respectively.

1) Inverted Residual Block: As shown in Fig. 2, the structure of each inverted residual block can be divided into three parts, expansion, depthwise convolution, and projection. Depthwise convolution presents excellent performance in high-dimensional feature space, whereas it is not available to change the number of channels. Thus, pointwise convolution is employed in expansion layer to increase the number of channels and in projection layer to decrease it. Each convolutional layer is followed by a batch normalization. In addition, a SiLU function is performed following the batch normalization in expansion and depthwise convolution. Finally, the result is fused with the original input feature by elementwise addition.

2) TAE-MobileVIT Block: The core idea of MobileVIT is to learn global dependencies with transformers as convolutions. The architecture of our TAE-MobileVIT block is shown in Fig. 3. First, bitemporal feature maps are transformed and divided into high-dimensional patches with spatial information through local representations. Given input tensors $F_{As}, F_{Bs} \in \mathbb{R}^{H\times W\times C}(s = 16,32$ denotes the size of tensors), an $n \times n$ standard convolutional layer is employed to encode the information of position. The following pointwise convolution projects the tensors into high-dimensional $\mathbb{R}^{H\times W\times d}$ by learning linear combinations of the input channels. Here, *C, H,* and *W* denote the channels, height, and width of the tensor, respectively, while *d* denotes the channels of the high-dimensional space. Natural language processing splits the input sentence into some words or phrases and represents each one with a token vector. Likewise, vision transformer [37] splits the input image into several patches. As is shown in Fig. 3, pixels in the same relative position of each patch correspond to one token vector. Therefore, high-dimensional tensors are unfolded into bitemporal token sets $T_{As}, T_{Bs} \in \mathbb{R}^{N\times P\times d}$. Here, $N = hw$ denotes the number of pixels in the patch with height h and width w, and $P = \frac{HW}{N}$ denotes the number of patches. N is set to 4 according to the parametric analysis in Section V. Specially, T_{As} and T_{Bs} are concatenated into $T_{ABs} \in \mathbb{R}^{N\times 2P\times d}$. For each $m \in \{1, \ldots, N\}$, interpatch relations are encoded by applying transformer to obtain $G_{ABs} \in \mathbb{R}^{N\times 2P\times d}$ as

$$G_{ABs}(m) = \text{Transformer}\left(T_{ABs}(m)\right),\ 1 \le m \le N. \quad (1)$$

Each patch contains pixels from different spatial-temporal domains, it follows that transformer is available to enhance

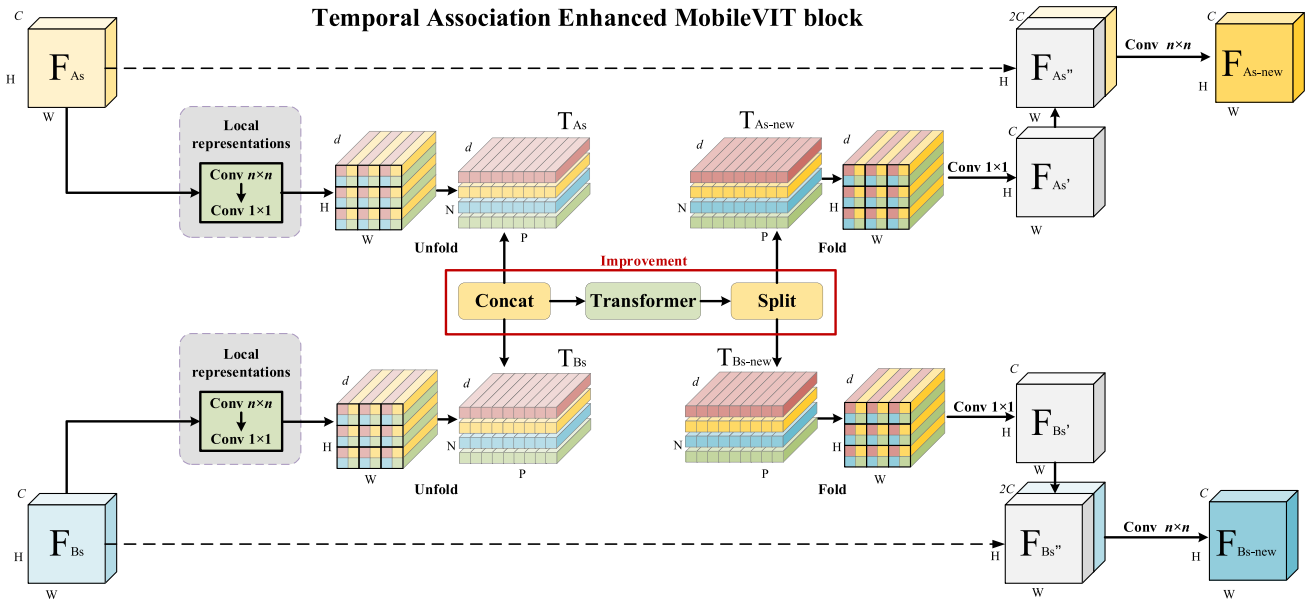

Fig. 3. Architecture of the TAE-MobileVIT block.

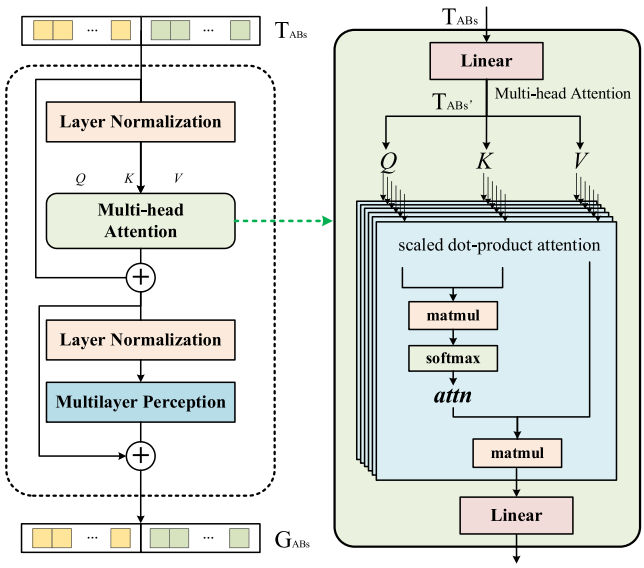

Fig. 4. Structure of the transformer and the multihead attention block.

the spatial-temporal association. G_{ABs} is split into context-rich $T_{As-\text{new}}, T_{Bs-\text{new}} \in \mathbb{R}^{N\times P\times d}$ and folded into $F'_{As}, F'_{Bs} \in \mathbb{R}^{H\times W\times d}$, which are subsequently projected to lower dimensional space via a pointwise convolution. F'_{As} is concatenated with the original input F_{As} to obtain F''_{As}, and F'_{Bs} is processed in the same way. An $n \times n$ convolution is then applied to F''_{As} and F''_{Bs} to obtain the output $F_{As-\text{new}}, F_{Bs-\text{new}} \in \mathbb{R}^{H\times W\times C}$.

The structure of the transformer in our TAE-MobileVIT block is shown in Fig. 4. It consists of layer normalization and two sublayers, multihead attention block and multilayer perception block. After obtaining the concatenated token set T_{ABs}, a layer normalization is applied before and after the multihead attention block. Then, a multilayer perception block is employed. In addition, a residual connection is employed after each sublayer.

In multihead attention block shown in Fig. 4, the input token set $T_{ABs} \in \mathbb{R}^{N\times 2P\times d}$ is first expanded into high-dimensional T'_{ABs} by a linear transformation as

$$T'_{ABs} = T_{ABs}W^{I}, T'_{ABs} \in \mathbb{R}^{N\times 2P\times(m\times l\times 3)} \quad (2)$$

where W^{I} is the weight of the embedding layer, m is the number of heads in multihead attention, and l is the dimension for subsequent tensors. m and l are set for 4 and 8, respectively, the same with the original literature of MobileVIT. Then T'_{ABs} is forwarded into different heads of multihead attention. The parameters are not shared among the heads. As is shown in Fig. 4, each head of the multihead attention consists of linear layers and scale dot-product attention. T'_{ABs} is transformed into query ($Q \in \mathbb{R}^{m\times N\times 2P\times l)}$), key ($K \in \mathbb{R}^{m\times N\times 2P\times l)}$), and value ($V \in \mathbb{R}^{m\times N\times 2P\times l)}$) by linear layers as

$$Q,\ K,\ V = T'_{ABs}W^{Q}, T'_{ABs}W^{K}, T'_{ABs}W^{V} \quad (3)$$

where W^{Q}, W^{K}, and W^{V} denote the weights of the linear layers to obtain Q, K, V.

The structure of scale dot-product attention is depicted in Fig. 4. We employ the dot-product operation and Softmax activation to obtain the attention map which is used as the weight of V. The process of scale dot-product attention can be expressed as

$$\text{SDPA}\,(Q,\ K,\ V) = \text{Softmax}\left(\frac{QK^{T}}{\sqrt{d}}\right)V. \quad (4)$$

Then, the output of each head is concatenated and input into a linear layer

$$\text{head}_i = \text{SDPA}\left(T'_{ABs}\,W_i^{Q}, T'_{ABs}\,W_i^{K}, T'_{ABs}\,W_i^{V}\right),$$
$$i \in \{1,\ \ldots,\ m\} \quad (5)$$

$$\text{MHA}\,(Q,\ K,\ V) = \text{Concat}\,(\text{head}_1, \ldots, \text{head}_m)W^{O} \quad (6)$$

where ${W_i}^Q$, ${W_i}^K$, and ${W_i}^V$ denote weights of the linear layers of the ith head to map Q, K, and V. ${W}^O$ denotes the weight of the last linear layer in multihead attention block.

The multilayer perception block, which contains two linear layers with a SiLU function inside, is applied to further transform the learning tokens of multihead attention. Then, the output token set is split into context-rich token sets $T_{As-\text{new}}$ and $T_{Bs-\text{new}}$.

B. SAE Multiscale Feature Difference Aggregator

As shown in Fig. 1, a SAE multiscale feature difference aggregator is used to fuse multiscale feature maps from the feature extractor. First, subtraction and absolute operations are employed on bitemporal feature maps to obtain feature difference maps of various scales $D_s \in \mathbb{R}^{H_s \times W_s \times C_s}$. Here, H_s and W_s ($s =$ 16,32,64) denote the size of feature difference maps, and C_s denotes the dimension of the corresponding size. Specifically, $H_s = W_s = s$, C_{16}= 128, C_{32}= 96, C_{64}= 64. Then, a pointwise convolution is used to transform each C_s to C_0= 64 in order to obtain the feature difference maps $D'_s \in \mathbb{R}^{H_s \times W_s \times C_0}$. Second, D'_{16} is enhanced through a convolutional block attention module (CBAM) [38] and upsampled to $U_s \in \mathbb{R}^{H_s \times W_s \times C_0}$ ($s =$ 32,64). Third, D'_{32} and D'_{64} are concatenated with U_{32} and U_{64}, respectively. Then, the concatenated tensors are input into a CBAM separately to obtain $M_s \in \mathbb{R}^{H_s \times W_s \times 2C_0}$ ($s =$ 32,64), which focus more on regions of interest. Finally, $M_{32}$is upsampled to $U'_{64} \in \mathbb{R}^{H_s \times W_s \times 2C_0}$($s = 64$) and concatenated with M_{64} to obtain $X \in \mathbb{R}^{H_s \times W_s \times 4C_0}$ ($s = 64$).

CBAM is a light-weight attention module, which concentrates on information of interest by adjusting the network weights adaptively. It contains two separate modules, channel attention module (CAM) and spatial attention module (SAM). In CAM, we first apply a global max pooling and a global average pooling based on the height and weight to the input feature F, respectively. The feature maps are severally input into a weight shared two-layer multilayer perception with a rectified linear unit function [39]. Then, an elementwise addition operation and a SiLU function are employed to obtain the channel attention feature M_C

$$\begin{aligned} M_C(F) &= \sigma\left(\text{MLP}\left(\text{AvgPool}\left(F\right)\right) + \text{MLP}\left(\text{MaxPool}\left(F\right)\right)\right) \\ &= \sigma\left(W_1\left(W_0\left(F^c_{\text{avg}}\right)\right) + W_1\left(W_0\left(F^c_{\text{max}}\right)\right)\right) \end{aligned} \tag{7}$$

where $\sigma(.)$ denotes the sigmoid function. W_1 and W_0 denote the shared MLP weights. The channel-refined feature F' can be denoted as

$$F' = M_C\ (F) \otimes F \tag{8}$$

where $\otimes$ denotes the elementwise multiplication.

The channel-refined feature F' can be input into SAM and enhanced with another global max pooling and global average pooling based on the channel. Then, we concatenate the feature maps and obtained the spatial attention feature M_S via a 7 × 7 convolution and a sigmoid attention

$$\begin{aligned} M_S(F) &= \sigma\left(f^{7\times7}\left(\left[\text{AvgPool}\left(F\right);\ \text{MaxPool}\left(F\right)\right]\right)\right) \\ &= \sigma\left(f^{7\times7}\left(\left[F^s_{\text{avg}}; F^s_{\text{max}}\right]\right)\right) \end{aligned} \tag{9}$$

where $\sigma(.)$ denotes the sigmoid function and $f^{7\times7}(.)$ denotes the size of the convolution kernel.

The output feature F'' can be denoted as

$$F'' = M_S(F') \otimes F' \tag{10}$$

where $\otimes$ denotes the elementwise multiplication.

C. Prediction Head

A context-rich feature difference map is generated benefiting from the feature extractor enhanced by temporal association and the SAE multiscale feature difference aggregator. Hence, a simple and shallow FCN is used in STAE-MobileVIT for change discrimination. In the prediction head, we apply a pointwise convolution on the concatenated feature map $X \in \mathbb{R}^{64\times64\times256}$. Then, X is upsampled to $X' \in \mathbb{R}^{256\times256\times32}$. The classifier employs two 3 × 3 convolutional layers with batch normalization and a Softmax function pixel-wisely operated on the channel dimension. Finally, the predicted change probability map $P \in \mathbb{R}^{256\times256\times2}$ can be expressed as following:

$$P = \sigma\left(\text{g}\left(X'\right)\right) \tag{11}$$

where X' denotes the upsampled feature difference maps, g(.) denotes the convolution, and $\sigma(.)$ denotes the Softmax function.

III. Experimental Setup

A. BSL-CD Dataset

For the lack of binary change detection dataset designed for BSL, we created a small dataset BSL-CD to investigate the feasibility of applying deep learning for BSL change detection. BSL-CD contains 1083 pairs of remote sensing images with high-spatial resolution of 0.8 m. The samples are randomly split into three parts: 779, 195, and 109 pairs for training, validation, and testing, respectively. The ratio of changed BSL pixels to other pixels is nearly 1:4 in BSL-CD and each of the abovementioned three sets. The bi-temporal images in BSL-CD were collected by the BJ-2 satellite in Daxing District, Beijing, China, in July and September 2020 as well as in January and May 2021. Each sample group consists of bitemporal images with the size of 256 × 256 pixels and a corresponding binary label indicating BSL change. As shown in Fig. 5, the main types of changes annotated in BSL-CD include buildings, vegetation, dust-proof nets, and construction sites. In practical application, we expect the extraction results to be entire pieces but not fragmented pixels. For this reason, during the process of labelling the reference images, we adhere the majority principle, ensuring that the classification of the complete pieces is not influenced by fragmented, heterogeneous pixels occupying less than 5% of the total area.

B. Comparison With Other Public Methods

In order to validate the results of STAE-MobileVIT, we made a comparison with seven methods, including three classical FCN-based models, a UNet-based model, a multiscale supervision model, a spatial-temporal attention-based model, and a traditional transformer-based model. The following change

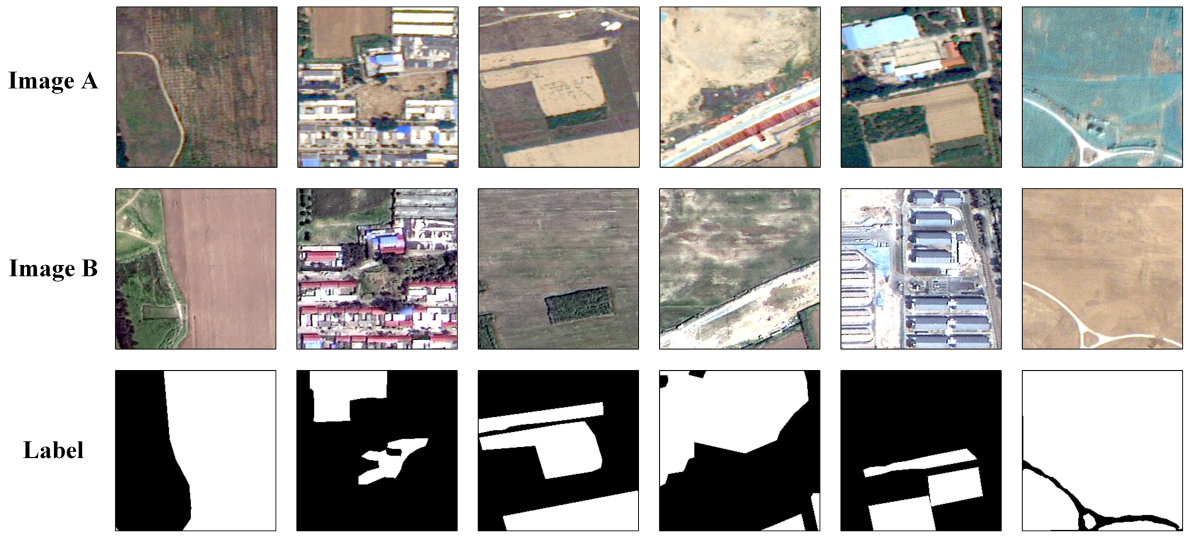

Fig. 5. Examples of samples in BSL-CD. (Image A denotes the previous image and Image B denotes the subsequent image.).

detection models were implemented using their public codes and hyperparameters were set as consistent as possible with the original literature.

1) FC-EF [15] is an image-level fusion method based on FCN, where the concatenated channels of bitemporal images are regarded as a single input.
2) FC-Siam-diff [15] is a feature-level fusion method based on FCN, where bitemporal features are extracted by Siamese networks, and the change information is then derived from the feature difference maps.
3) FC-Siam-conc [15] is a feature-level fusion method based on FCN, in which Siamese networks are employed to extract bitemporal features and the feature concatenation is used to obtain the change information.
4) CDNet [16] is an image-level fusion method based on UNet, where the concatenated channels of bitemporal images are regarded as a single input.
5) Deeply supervised image fusion network (DSIFN) [17] can fuse multiscale features and feature difference maps through the spatial and channel attention mechanism, and it applies deep supervision (i.e., computing supervised loss at each level of the decoder) to train the intermediate layers.
6) Spatial-temporal attention neural network (STANet) [18] is based on Siamese FCN, which can extract more discriminative change information due to the integration of the spatial-temporal attention mechanism.
7) Bitemporal image transformer (BIT) [25] is a transformer-based method, which incorporates the Siamese tokenizers and transformer into conventional Siamese change detection networks in order to capture rich contexts.

C. Accuracy Evaluation Metrics

In order to evaluate the predicted results, we use three common metrics, overall accuracy (OA), intersection over union (IoU), and F1-score. They can be defined as follows:

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{12}$$

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{13}$$

$$\mathrm{F1} = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} \tag{14}$$

where TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively. We consider the scientific metric F1-score, which combines recall and precision, as the main assessment index. Recall and precision can be defined as follows:

$$\mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{15}$$

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}. \tag{16}$$

D. Implementation Details

All the experiments involved were implemented on PyTorch and models were trained using Intel Core i7-8700K processor, 32 GB memory, NVIDIA Geforce RTX 2080 Ti graphics card (11 GB). We applied normal data augmentation to the input image patches, including flip, rescale, crop, and Gaussian blur. STAE-MobileVIT was trained with the batch size of 8 using a stochastic gradient descent optimizer with momentum of 0.99 and weight decay of 0.0005. The learning rate was initialized to 0.01 and linearly decayed to 0 during the 200 epochs. The cross-entropy loss was minimized to optimize the network parameters.

IV. Results and Analysis

Table I reports the F1, IoU, and OA results repeated 3 times on BSL-CD. The quantitative results show that STAE-MobileVIT consistently outperforms the other methods with a significant margin. STANet gains the lowest and the least

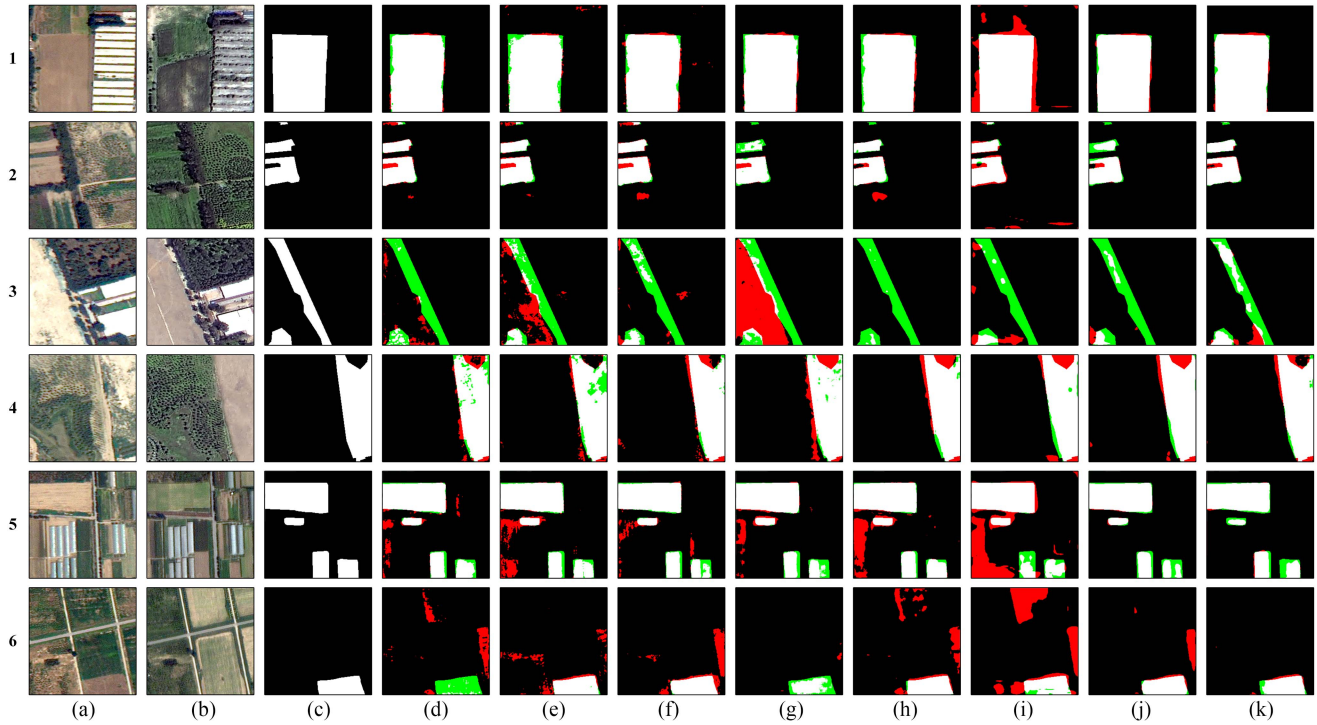

Fig. 6. Visualization results of comparison to seven common methods on BSL-CD. Different colors are used for better view, i.e., white for TP, black for TN, red for FP, and green for FN. (a) Image A. (b) Image B. (c) Label. (d) FC-EF. (e) FC-Siam-conc. (f) FC-Siam-diff. (g) CDNet. (h) DSIFN. (i) STANet. (j) BIT. (k) STAE-MobileVIT. (Image A denotes the previous image and Image B denotes the subsequent image.).

TABLE I
COMPARISON RESULTS ON BSL-CD

Method	F1(%)	IoU(%)	OA(%)
FC-EF	80.96±0.51	68.02±0.72	91.57±0.27
FC-Siam-conc	78.17±0.23	64.16±0.31	90.07±0.18
FC-Siam-diff	78.98±0.47	65.27±0.65	90.89±0.22
CDNet	78.19±0.14	64.19±0.19	90.55±0.03
DSIFN	79.07±0.71	65.39±0.97	90.81±0.39
STANet	77.99±1.14	63.93±1.53	90.42±0.50
BIT	78.30±0.65	64.34±0.88	90.83±0.29
STAE-MobileVIT	**84.44±0.43**	**73.07±0.65**	**93.01±0.17**

The optimal results are marked in bold.

stable F1 (77.99 $\pm$ 1.14%) and IoU (63.93 $\pm$ 1.53%) among all the methods, while FC-Siam-conc and CDNet are slightly higher but much steadier. FC-EF achieves 80.96 $\pm$ 0.51% of F1 (68.02 $\pm$ 0.72% of IoU and 91.57 $\pm$ 0.27% of OA), which is the best result among the seven methods. The performance of FC-Siam-diff, DSIFN, and BIT are moderate between CDNet and FC-EF with 78.98 $\pm$ 0.47%, 79.07 $\pm$ 0.71% and 78.30 $\pm$ 0.65% of F1, respectively. On the whole, STAE-MobileVIT reaches the optimal and relatively stable F1, IoU and OA (84.44 $\pm$ 0.43%, 73.07 $\pm$ 0.65% and 93.01 $\pm$ 0.17%, respectively), which are 3.48, 5.05, and 1.44 percent higher than the second-ranked FC-EF.

The comparison of STAE-MobileVIT with the seven methods on BSL-CD is visually depicted in Fig. 6. For a better view, different colors are used to denote TP (white), TN (black), FP (red), and FN (green). It can be observed that changes between BSL and lush vegetation or buildings can be extracted accurately by most of the abovementioned methods. However, certain types of BSL change remain prone to be misclassified, yet the proposed method achieves better results than others. First, STAE-MobileVIT can reduce the impacts of pseudochanges, because the bitemporal association is enhanced by TAE-MobileVIT block. For example, sparse grassland in row 5 and 6 (red FP), which is mistaken as BSL in results of the seven methods, can be classified correctly by STAE-MobileVIT. The change detection of BSL is various from other land use/cover types like buildings, ships, and aircrafts, which are easy to confirm whether there is or not. It is significantly affected by pseudochanges, because sparse grassland has similar color behaviors as BSL. The proposed TAE-MobileVIT block is available to overcome the influence of bitemporal pseudo changes effectively. Second, STAE-MobileVIT can also well-handle the fine-grained alteration of BSL. An example of the distinctly discriminated grained vegetation and dust-proof nets from BSL is shown in row 2 (red FP) and 3 (green FN). Compared with FC-EF, FC-Siam-conc, and FC-Siam-diff, more details of changes are rendered in the results of STAE-MobileVIT, since the SAE multiscale feature difference aggregator integrates multiscale features and allows the shallow features to provide detailed information. Similarly, DSIFN and STANet also perform well in fine-grained change detection on account of their multiscale features fusion strategy. Third, STAE-MobileVIT perform well in extracting pieces of land in regular shape and obtaining entire results with minimal noise. This suggests that BIT and STAE-MobileVIT benefit from transformer, which is available to model global dependencies to alleviate the loss of long-range information. For instance, from row 1, 4, 5, and 6, BIT and STAE-MobileVIT detect complete pieces of land with regular boundary, while other methods fail

TABLE II
ABLATION EXPERIMENTS RESULTS ON BSL-CD

Method	F1(%)	IoU(%)	OA(%)
Base	81.41±0.77	68.66±1.10	91.99±0.29
+SAE	83.23±0.21	71.28±0.30	92.61±0.05
+TAE	82.42±0.57	70.09±0.82	92.46±0.32
STAE-MobileVIT	**84.44±0.43**	**73.07±0.65**	**93.01±0.17**
STAE-MobileVIT-xs	83.63±0.98	71.87±1.44	92.79±0.29
STAE-MobileVIT-xxs	83.34±0.86	71.45±1.27	92.82±0.28

The optimal results are marked in bold.

TABLE III
EFFICIENCY OF VARIOUS MODELS ON BSL-CD

Method	FLOPs(G)	Params(M)	Time(h)
FC-EF	28.62	1.35	1.66±0.05
FC-Siam-conc	42.65	1.54	1.87±0.03
FC-Siam-diff	37.81	1.35	1.78±0.03
CDNet	187.79	1.43	2.45±0.05
DSIFN	658.05	35.73	3.69±0.47
STANet	100.48	16.89	4.47±0.24
BIT	85.07	3.50	0.87±0.18
STAE-MobileVIT	40.39	2.80	0.66±0.09
STAE-MobileVIT-xs	27.36	1.34	0.75±0.06
STAE-MobileVIT-xxs	**16.92**	**0.73**	**0.62±0.09**

The optimal results are marked in bold.

due to the limited receptive fields of CNN. However, the linear feature from row 3 (green FN) requires more attention, and the connectivity of results needs to be improved. To summarize, owing to the advantages of the transformer structure, the SAE multiscale feature difference aggregator and the TAE-MobileVIT block, STAE-MobileVIT surpasses the comparative methods in preserving details, reducing misclassification and mitigating the effects of pseudochanges. Nevertheless, the proposed method leaves the extraction of linear features and connectivity to be improved.

V. DISCUSSION

A. Ablation Study

The ablation study on BSL-CD is performed to further verify the effectiveness of SAE denoting the SAE multiscale feature difference aggregator and TAE denoting the TAE-MobileVIT block, and both are integrated in STAE-MobileVIT. The "Base" model employs the Siamese MobileVIT as the encoder and a two-layer FCN as the decoder for change detection. "+SAE" represents the "Base" model with SAE multiscale feature difference aggregator and "+TAE" represents the "Base" model with the TAE-MobileVIT block. In addition to modifying from MobileVIT, we also use lighter-weight MobileVIT-xs and MobileVIT-xxs as the basic network, which have less channels and dimensions. The results of the ablation study repeated 3 times are given in Table II. We can find consistent and significant drops in F1, IoU, and OA when removing SAE, TAE, or both of them from STAE-MobileVIT, which indicates the vital importance of SAE and TAE. Compared with the "Base" model with F1 of 81.07%, the F1 of "+SAE" and "+TAE" are improved by 1.95 and 1.44 percent, respectively. The results with F1, IoU, and OA for STAE-MobileVIT reaches 83.98%, 72.39%, and 92.93% respectively, which proves the validity of the integration of SAE and TAE. The decline of standard deviation in "+SAE," "+TAE," and STAE-MobileVIT signifies the excellent stability of modified models. Besides, the F1 and IoU slightly decrease when using lighter-weight STAE-MobileVIT-xs and STAE-MobileVIT-xxs, while the OA is almost unchanged. This suggests that fewer parameters lead to higher volatility in results, albeit with a slight decrease in accuracy.

Fig. 7 visualizes the comparison of the ablation results. It can be seen from row 2, 3, 4, and 5 that compared with the "Base" model, FP (red), and FN (green) can be better avoided in the results of "+SAE" and "+TAE". More spatial details are kept in "+SAE," because SAE aggregates semantic characteristics and detailed information simultaneously, which is effective to modify the results of BSL change detection. From row 3 and 5, the fragmented misclassifications in "+TAE" are fewer than that in "+SAE," since the transformer structure in TAE models global dependencies to render more complete results. TAE also enhances the association of bitemporal images in order to exclude pseudochanges. From the visualized examples, STAE-MobileVIT gains the best results, while the lighter-weight STAE-MobileVIT-xs and STAE-MobileVIT-xxs also perform better than "+SAE" and "+TAE" results.

B. Efficiency and Effectiveness of Models

In order to further explore the efficiency and effectiveness of the models, we use three common metrics, floating-point operations (FLOPs), size of parameters (Params), and training time (Time) to measure the computational costs and assess the efficiency. FLOPs whose unit is 10^9 (G) measure the computational complexity via the computing time of multiplication and addition operations for a pair of input images with the size of $256 \times 256 \times 3$. Params whose unit is 10^6 (M) measure the number of parameters to be learned during training, which is related to the complexity of the model. Time whose unit is hours (h) measures the computational cost by recording the average training time per 100 epochs repeated three times.

The FLOPs, Params, and Time of all the methods are illustrated in Table III. Compared with the seven methods, the FLOPs of STAE-MobileVIT are slightly more than FC-EF, close to FC-Siam-conc and FC-Siam-diff, and notably less than the other models. The Params of STAE-MobileVIT are observably less than DSIFN and STANet, and close to others. The training time of STAE-MobileVIT is slightly less than BIT and significantly less than the other six methods. In order to provide more efficient and effective models, STAE- MobileVIT-xs and STAE-MobileVIT-xxs are performed with the least FLOPs, Params, and training time. The performances are even better than the seven methods and nearly as excellent as STAE-MobileVIT from the results in the above Tables I and II. Accordingly, STAE-MobileVIT is available to balance the accuracy and computational costs.

C. Parameter Analysis

1) Patch Size: Our institution is to unfold high-dimensional tensors into several token vectors, and encode bitemporal token sets to enhance the spatial-temporal association. The number of

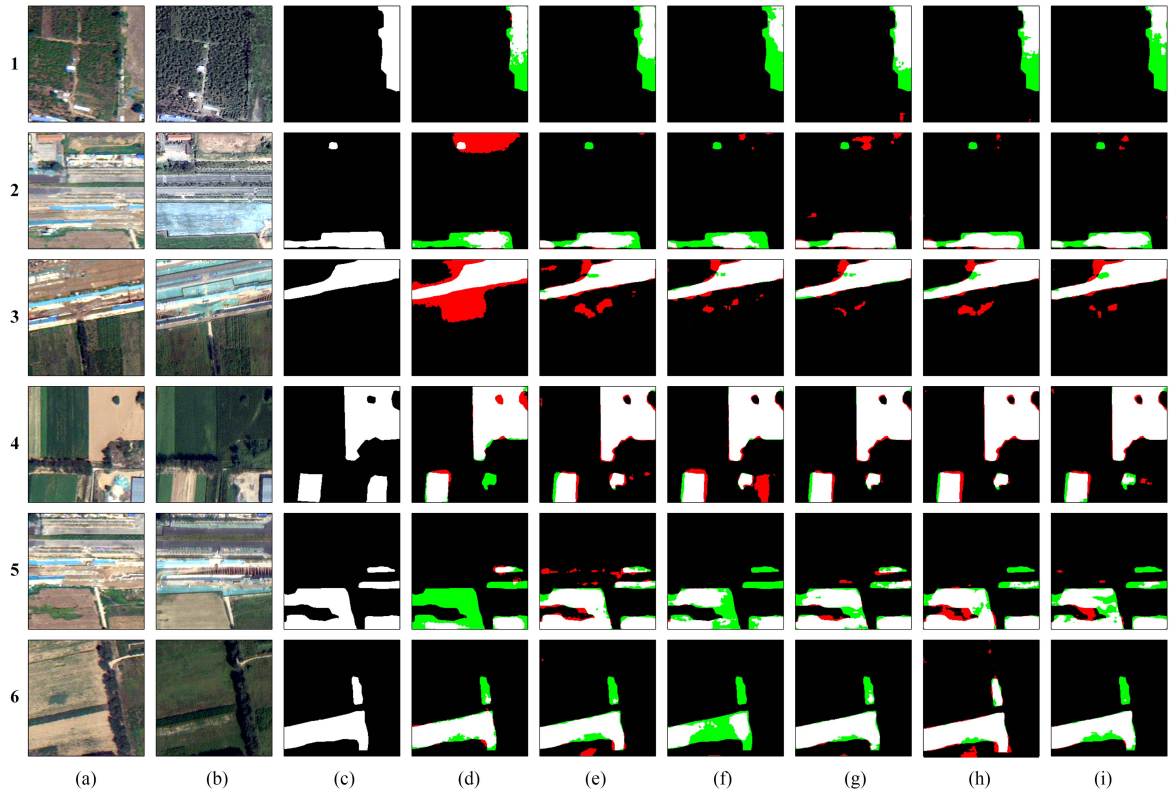

Fig. 7. Visualization results of ablation study on BSL-CD. Different colors are used for better view, i.e., white for TP, black for TN, red for FP, and green for FN. (a) Image A. (b) Image B. (c) Label. (d) Base. (e) +SAE. (f) +TAE. (g) STAE-MobileVIT. (h) STAE-MobileVIT-xs. (i) STAE-MobileVIT-xxs. (Image A denotes the previous image and Image B denotes the subsequent image.).

TABLE IV
EFFECT OF PATCH SIZE

Patch	F1(%)	IoU(%)	OA(%)	FLOPs(G)
4	**84.44**	**73.07**	**93.01**	40.39
16	83.88	72.23	92.88	40.32
64	83.41	71.55	92.62	**40.29**

The optimal results are marked in bold.

TABLE V
EFFECT OF DEPTH OF TRANSFORMER

D16	D32	F1(%)	IoU(%)	OA(%)	FLOPs(G)
2	2	82.90	70.80	92.58	**37.76**
4	2	**84.44**	**73.07**	93.01	40.39
8	2	82.88	70.78	92.59	45.67
4	4	83.62	71.86	92.65	43.86
4	8	84.30	72.87	**93.06**	50.78

The optimal results are marked in bold.

token vectors is an important hyperparameter depending on the number of pixels in each patch. We tested different patch size $N \in \{4, 16, 64\}$ to analyze its effect on the performance of STAE-MobileVIT. Table IV reports the average F1, IoU, and OA results repeated three times on BSL-CD. There is a significant decline in F1 when increasing the patch size from 4 to 64. It indicates that compact token sets are sufficient to denote fine-grained global features and redundant token sets may hinder the performance of the model. On the other hand, larger patch sizes result in shorter token length and relatively less computational costs. As the STAE-MobileVIT is a light-weight model, the decrease of FLOPs is rather obscure. Therefore, we set the patch size to 4.

2) Depth of Transformer: The depth of transformer is also one important hyperparameter influencing the results of the extraction. We tested different configurations of STAE-MobileVIT containing various numbers of transformer layers in TAE-MobileVIT blocks. The depths of transformer in 16×16 and 32×32 feature layers are set to $\mathrm{D}_{16}, \mathrm{D}_{32} \in \{2, 4, 8\}$. As shown in Table V, too large or small D_{16} decreases the precision of the results. It indicates that a limited number of transformer layers is insufficient to encode tokens with rich semantic information, while an excessive number of layers is also not beneficial. The moderate depth of D_{16} proves to be more effective. Besides, Table V shows no significant improvements in F1 when increasing the depth of D_{32}. This suggests that relationships between bitemporal tokens in shallow layers can be well-learned by a shallow transformer. Table V also shows that FLOPs is positively correlated with the depths of the transformer. Accordingly, for the tradeoff between efficiency and precision, the depth of transformer $\mathrm{D}_{16}, \mathrm{D}_{32}$ are set to 4 and 2, respectively.

D. Effect of Dataset Size

To further explore the effectiveness of STAE-MobileVIT on small datasets, we conducted experiments using various numbers of samples for training and validation of the model. Table VI presents the experimental results obtained from the same testing dataset. It is evident that the proposed method

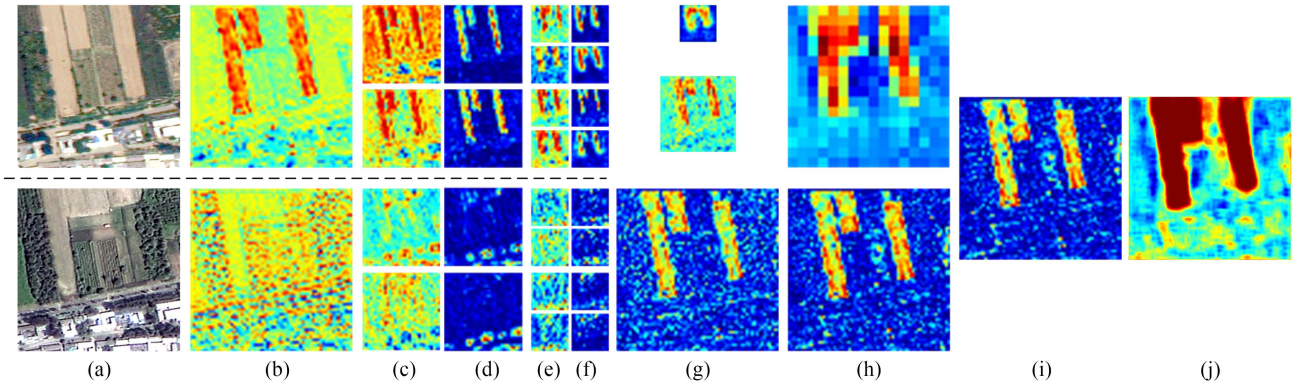

Fig. 8. Example of network visualization. (Red denotes higher attention values and blue denotes lower values.) (a) Input bitemporal images. (b) F_{A64}, F_{B64}. (c) F_{A32}, F_{B32}. (d) Temporal association enhanced $F_{A32\text{-new}}$, $F_{B32\text{-new}}$. (e) F_{A16}, F_{B16}. (f) Temporal association enhanced $F_{A16\text{-new}}$, $F_{B16\text{-new}}$. (g) D_{16}, D_{32}, and D_{64}. (h) Upsampled M_{32} and M_{64}. (i) Fusion of multiscale feature difference maps. (j) Change probability map.

TABLE VI
EFFECT OF DATASET SIZE

Train	Val	F1(%)	IoU(%)	OA(%)
779	195	**84.44**	**73.07**	**93.01**
544	156	80.60	67.51	91.86
389	111	79.19	65.55	91.02

The optimal results are marked in bold.

exhibits exceptional performance when working with small datasets. For example, the F1, IoU, and OA of relatively small datasets of 500 and 800 samples for training are higher than 79.19, 65.55, and 91.02, demonstrating the effectiveness of STAE-MobileVIT on small datasets. As indicated in Table VI, a notable decrease in F1, IoU, and OA can be observed as the size of the training set decreases. Therefore, in terms of small datasets, a larger size of the training set can provide a greater amount of information regarding positive and negative samples, ultimately resulting in higher accuracy.

E. Network Visualization

To further understand the proposed model, an example is provided to visualize the feature maps at different stages of STAE-MobileVIT, as shown in Fig. 8. Given bitemporal images (a), a Siamese convolution and inverted residual blocks generate the feature maps with the size of 64 × 64 (b), 32 × 32 (c) and 16 × 16 (e). TAE-MobileVIT block enhances the bitemporal association of feature maps with the size of 32 × 32 (d) and 16 × 16 (f). Then, multiscale feature difference maps (g) of bitemporal images are obtained via subtraction and absolute operations. The upsampled M_{32} and M_{64} (h) are fused (i) and change probability map (j) is obtained through the classifier.

VI. CONCLUSION

Most BSL contributes to dust pollution and monitoring its change is of great significance for accurate environmental governance. We proposed an efficient CNN-transformer method STAE-MobileVIT and established a new dataset (BSL-CD) of high-resolution for BSL change detection. STAE-MobileVIT applies Siamese networks to extract features of bitemporal images, and it enhances the spatial-temporal association of bi-temporal features with TAE-MobileVIT block. Then, the multiscale feature difference maps are calculated and aggregated to further strengthen the spatial features. In the end, a simple FCN is employed as the classifier to predict the change detection results.

Extensive experiments on BSL-CD verify the capability of STAE-MobileVIT compared with six CNN-based methods and a transformer-based method. The proposed method exceeds the best results of the seven methods 3.48, 5.05, and 1.44 percent on F1, IoU, and OA. The ablation study validates the effectiveness of TAE and SAE. Specifically, TAE reduces the impacts of pseudochanges and misclassification by enhancing bitemporal association, and transformer can model global dependencies to alleviate the loss of long-range information. SAE ensures entire and detailed results via integrating multiscale feature difference maps. Moreover, the proposed model can be categorized based on the size into three variants: 1) the original STAE-MobileVIT, 2) the lighter-weight STAE-MobileVIT-xs, and 3) the even lighter STAE-MobileVIT-xxs. These three models with different sizes can be utilized for specific objectives, such as pursuing optimal precision and stability or striving for minimal computational complexity and costs. In conclusion, all the experimental results conducted on BSL-CD in this study demonstrate the efficiency and effectiveness of STAE-MobileVIT.

Nevertheless, there are still some improvements for the proposed method. For instance, to further modify the extraction of linear features and the connectivity of results, increased emphasis could be placed on boundary information. In addition, postprocessing techniques could be employed to optimize the final results.

ACKNOWLEDGMENT

The authors would like to thank to the people who helped in the acquisition of the satellite images for this article and also like to thank the handling editor and anonymous reviewers whose valuable and constructive comments greatly improved this article.

REFERENCES

[1] T. Li, Y. Feng, X. Bi, Y. Zhang, and J. Wu, "Main problems and refined solutions of urban fugitive dust pollution in China," *Environ. Sci.*, vol. 43, no. 3, pp. 1323–1331, 2022.

[2] H. Xu, "Dynamics of bare soil in A typical reddish soil loss region of Southern China: Changting County, Fujian Province," *Sci. Geographical Sinica*, vol. 33, no. 4, pp. 489–496, 2013.

[3] B. Chen and X. Zhou, "Explanation of current land use condition classification for national standard of the People's Republic of China," *J. Nat. Resour.*, vol. 22, no. 6, pp. 994–1003, 2007.

[4] Z. Wang, Y. Liu, Y. Ren, and H. Ma, "Object-level double constrained method for land cover change detection," *Sensors*, vol. 19, no. 1, p. 79, 2018.

[5] T. L. Sohl, "Change analysis in the United Arab Emirates: An investigation of techniques," *Photogrammetric Eng. Remote Sens.*, vol. 65, no. 4, pp. 475–484, 1999.

[6] D. Yuan and C. Elvidge, "NALC land cover change detection pilot study: Washington DC area experiments," *Remote Sens. Environ.*, vol. 66, no. 2, pp. 166–178, 1998.

[7] R. D. Johnson and E. Kasischke, "Change vector analysis: A technique for the multispectral monitoring of land cover and condition," *Int. J. Remote Sens.*, vol. 19, no. 3, pp. 411–426, 1998.

[8] 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.

[9] K. C. Seto, C. Woodcock, C. Song, X. Huang, J. Lu, and R. Kaufmann, "Monitoring land-use change in the Pearl River Delta using Landsat TM," *Int. J. Remote Sens.*, vol. 23, no. 10, pp. 1985–2004, 2002.

[10] J. Rogan, J. Franklin, and D. A. Roberts, "A comparison of methods for monitoring multitemporal vegetation change using Thematic Mapper imagery," *Remote Sens. Environ.*, vol. 80, no. 1, pp. 143–156, 2002.

[11] X. Zhang, D. Li, J. Gong, and Q. Qin, "A change detection method of integrating remote sensing and GIS," *Geomatics Inf. Sci. Wuhan Univ.*, vol. 31, no. 3, pp. 266–269, 2006.

[12] Y. Zhigao, Q. Qianqing, and Z. Qifeng, "Change detection in high spatial resolution images based on support vector machine," in *Proc. IEEE Int. Symp. Geosci. Remote Sens.*, 2006, pp. 225–228.

[13] W. Feng, H. Sui, J. Xu, K. Sun, and . Huang, "Change detection method for high resolution remote sensing images using random forest," *Acta Geodaetica Cartographica Sinica*, vol. 46, no. 11, pp. 1880–1890, 2017.

[14] C. Benedek and T. Szirányi, "Change detection in optical aerial images by a multilayer conditional mixed Markov model," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 10, pp. 3416–3430, Oct. 2009.

[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 Obs. Remote Sens.*, vol. 14, pp. 1194–1206, 2020.

[16] P. F. Alcantarilla, S. Stent, G. Ros, R. Arroyo, and R. Gherardi, "Street-view change detection with deconvolutional networks," *Auton. Robots*, vol. 42, pp. 1301–1322, 2018.

[17] C. Zhang et al., "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 166, pp. 183–200, 2020.

[18] 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, p. 1662, 2020.

[19] R. C. Daudt, B. L. Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proc. IEEE 25th Int. Conf. Image Process.*, 2018, pp. 4063–4067.

[20] C. Zhang, "Research on change detection in high resolution remote sensing images based on high level feature analysis," Ph.D. dissertation, Wuhan University, Wuhan, China, 2020.

[21] J. Geng, X. Ma, X. Zhou, and H. Wang, "Saliency-guided deep neural networks for SAR image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7365–7377, Oct. 2019.

[22] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 5998–6008, 2017.

[23] J. Lanchantin, T. Wang, V. Ordonez, and Y. Qi, "General multi-label image classification with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16478–16488.

[24] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 2021, pp. 6881–6890.

[25] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5920416.

[26] W. G. C. Bandara and V. M. Patel, "A transformer-based siamese network for change detection," in *Proc. IGARSS-IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 207–210.

[27] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Comput. Vision - ECCV*, 2018, pp. 833–851, doi: 10.1007/978-3-030-01234-2_49.

[28] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6848–6856.

[29] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:170404861*.

[30] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," 2021, *arXiv:211002178*.

[31] M. Liu, Z. Chai, H. Deng, and R. Liu, "A CNN-transformer network with multiscale context aggregation for fine-grained cropland change detection," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 15, pp. 4297–4306, 2022.

[32] K. Yang et al., "Asymmetric Siamese networks for semantic change detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5609818.

[33] S. Tian, A. Ma, Z. Zheng, and Y. Zhong, "Hi-UCD: A large-scale dataset for urban semantic change detection in remote sensing imagery," 2020, *arXiv:201103247*.

[34] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," JMLR.org, 2015.

[36] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural Netw.*, vol. 107, no. SI, pp. 3–11, Nov. 2018, doi: 10.1016/j.neunet.2017.12.012.

[37] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:201011929*.

[38] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[39] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist., JMLR Workshop Conf. Proc.*, 2011, pp. 315–323.

Sasha Wu received the B.S. degree in geographic information science from Sun-Yat-sen University, Guangzhou, China, in 2021. She is currently working toward the M.S. degree in cartography and geographic information system with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

Her research interests include intelligent interpretation of remote sensing images, deep learning, and change detection.

Yalan Liu received the Ph.D. degree in cartography and remote sensing from the University of Chinese Academy of Sciences (CAS), Beijing, China in 2004.

She is currently a Professor with the Aerospace Information Research Institute (originally Institute of Remote Sensing and Digital Earth), CAS. She was an Assistant Professor with the Department of GIS, Institute of Remote Sensing Applications, CAS, from 1996 to 2001, an Associate Professor from 2001 to 2008, and a Professor from 2008 to 2012. She has authored or coauthored more than 100 scientific articles in refereed journals and proceedings. Her research interests include the environmental applications of remote sensing and the integration of spatial information for smart cities and sustainable development.

Dr. Liu was Guest Editor and reviewer of several international journals such as *ISPRS International Journal of Geo-Information, Remote Sensing,* and *Land*.

Shufu Liu received the B.S. degree in cartography and geographic information system from Wuhan University, Wuhan, China, in 2007, and the Ph.D. degree in cartography and geographic information system from the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China, in 2013.

He is currently an Associate Professor with the Aerospace Information Research Institute (originally Institute of Remote Sensing and Digital Earth), Chinese Academy of Sciences, Beijing, China. His research interests include remote sensing image processing and ecological environment remote sensing application.

Dacheng Wang received the M.S. and Ph.D. degrees in agriculture remote sensing and information technology from Zhejiang University, Hangzhou, China, in 2004 and 2012, respectively.

He is currently a Senior Engineer with the Aerospace Information Research Institute (originally Institute of Remote Sensing and Digital Earth), Chinese Academy of Sciences, Beijing, China. In recent years, he has been devoted to the research of remote sensing application of Low Carbon Environment in Daxing Functional area of Beijing. His research interests include geospatial analysis and smart city project.

Linjun Yu received the Ph.D. degree in land use and information technology from China Agricultural University, Beijing, China, in 2012.

He is currently a Research Associate with the Aerospace Information Research institute (originally Institute of Remote Sensing and Digital Earth), Chinese Academy of Sciences, Beijing, China. He was a Visiting Scholar with the Department of Architecture and Regional Planning, University of Florida, USA, from 2009 to 2011. His research interests include spatial pattern simulation and analysis, land use planning, and sustainable development.

Yuhuan Ren received the Ph.D. degree in cartography and geography information system from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China, in 2010.

She is currently an Associate Professor with the Aerospace Information Research Institute (originally Institute of Remote Sensing and Digital Earth), Chinese Academy of Sciences, Beijing, China. She was an Assistant Research Fellow with the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, from 2010 to 2018. Her research interests include information extraction from remote sensing images and its application in urban.