ABSTRACT Remote sensing-based change detection (CD) is a critical technique of detecting land surface changes in earth observation. Inspired by recent success of a lightweight CD network – 3M-CDNet, we implemented 9 meaningful modifications to it, for example, the incorporation of MHSA (Multi-Head Self-Attention). This elaborately designed model is termed as 3M-CDNet-V2. Its effectiveness and advantages were demonstrated on three engineering CD datasets, and experimental results indicated that: (1) relative to other state-of-the-art algorithms, the proposed model obtained very competitive or slight better performance on both visual comparison and quantitative metrics evaluation. (2) By applying a novel transfer learning strategy, 3M-CDNet-V2 could perform well on the small dataset. (3) The incorporation of MHSA brings a substantial accuracy improvement while moderately increasing the computational complexity. (4) The late fusion framework and deep supervision contribute most to the performance gain of 3M-CDNet-V2, while the introduction of low-level features to the classifier by skip connection could guide the model to focus on detailed spatial information such as small changes, narrow shaped objects, or accurate boundaries. We hope that our 3M-CDNet-V2 model helps in improving the understanding of network architecture design for CD.

INDEX TERMS Neural network, remote sensing, medium weight, change detection.

I. INTRODUCTION
Change detection (CD) based on remote sensing (RS) technology is widely used to discover and identify differences in ground objects using two or more images in the same geographical location [1]. The development of this technology has attracted the interest of many researchers in the RS community, it has been applied in many fields, such as illegal construction identification, resource surveys, disaster monitoring, urban planning, mine-environmental investigation [2], and so on. Benefiting from the rapid development of high spatial resolution and multi-temporal RS earth observation, high-resolution multispectral images have gradually become the primary data source for many RS applications, especially change detection [1], [3]. However, although the fine contextual information and complex spatial characteristics that high-resolution images convey offer rich spatial details for land use/cover change analysis, they face several serious challenges too [4]. Difficulties arise from the enhanced intra-class variabilities [5], the spatial displacement due to the parallax distortion of ground objects (especially for high-rise buildings) [6], the confused spectral features of different objects, complex scenarios, illumination, camera motion, shadows, misregistration error, and etc.

Due to the great advantages in deep feature representation and non-linear problem modeling [5], deep learning technology has opened up new opportunities for CD tasks to address the problems above-mentioned. For example, the CNN (i.e., Convolutional Neural Networks) architecture itself could achieve some degree of shift, scale, and distortion...
In addition, 3M-CDNet is incorporated with deformable convolutions (termed 3M-CDNet-V2) for change detection tasks using bitemporal high-resolution imagery. As its name implies, 3M-CDNet-V2 goes even deeper – half the size of the original network: 3M-CDNet, we made a series of modifications on the classical CDD and BCDD datasets [5]. Siamese CNN architectures have evolved to become increasingly powerful through more dense skip connections [5], larger scale [11], more sophisticated forms of convolution [14], semi-supervised architectures [12], [13], as well as hybrid loss function [15]. At the same time, the Transformer-based network architectures used for CD have also received increased attention [23], [24], [25], and they achieved competitive performances compared to their CNN counterparts. By leveraging the multi-head self-attention, Transformer has larger receptive fields and is notable for its use of attention to model long-range dependencies in the data [25]. On the contrary, most advanced CD pipelines using CNNs are still struggling to relate long-range concepts in space-time because of their inherent limitation to the size of the receptive field [23], [26]. Even so, Transformers have their inherent limitations, e.g., the low degree of computational and memory efficiency [7], the lack of convolution-like inductive biases [27], and so on.

Inasmuch, the models that interleave or combine CNN and Transformer in a “hybrid” way have fueled successful results [27] in a variety of computer vision tasks including object detection, image classification, and semantic segmentation, such as ConViT [28], TransCNN [29], ViT [30], CvT [7], and so on. However, there are very few studies examining the use of such hybrid methods in the realm of CD. In a more recent work [24], a transformer architecture is applied in conjunction with ResNet to enhance the feature representation while keeping the overall CNN-based feature extraction process in place. In this article, we will report a novel Siamese CNN-self attention combined architecture (termed 3M-CDNet-V2) for change detection tasks using bitemporal high-resolution imagery. As its name implies, 3M-CDNet-V2 is the updated version of 3M-CDNet [11] which is a lightweight CD network only involving about 3.12 M trainable parameters. 3M-CDNet was designed by reducing the width and depth of the backbone network, thus producing high-resolution feature maps, and it facilitates the detection of detailed spatial information (including small changed objects) with acceptable computational costs. In addition, 3M-CDNet is incorporated with deformable convolutions to promote the geometric transformation modeling ability. It makes a better trade-off between accuracy and inference speed compared with existing methods. However, in practical applications, we found that 3M-CDNet has some limitations as well, such as the vanishing-gradient problem, suboptimal improvement in model performance, difficulties in model training, limited power of deformable convolutions, and so on. We also found that these limitations could be minimized by a series of modifications, such as replacing several deformable convolutions in 3M-CDNet with Multi-Head Self-Attention (MHSA) layers that implement global self-attention over a 2D feature map [31]. These findings have motivated us to develop an updated version of 3M-CDNet by combining CNN and MHSA into a singular network.

In addition, the classic CNNs or CNN-like ViT networks often involve too many pooling or down-sampling operations, and the spatial details in deep features are lost. To solve this issue, the new generation of CNN models usually emphasizes the importance of retention and fusion of both the spatial detail information and semantic information, such as IFN [16], SNU-Net-CD [17], etc. Moreover, the classic ViT network [30] does not perform down-sampling operations, but it can enrich the high-level semantic information with detailed spatial features. On the other hand, if the training and inference of a model were directly based on high-resolution feature maps, the gaps between detailed spatial information and semantics could be bridged or at least narrowed. 3M-CDNet is such an example – the feature maps to be fed into the classifier are a quarter of the original image size, and 3M-CDNet-V2 goes even deeper – half the size of the original image.

As mentioned, to obtain better prediction accuracy, there is a trend to combine the strengths of CNNs and ViT (MHSA) [7]. There are three possible ways to do this: (1), most recent ViTs, such as CvT [7] and ConViT [28], adopt a hierarchical architecture, which is widely used in CNNs, e.g., AlexNet and ResNet. These models apply the global MHSA and its variants on the high-resolution tokens, thus bringing heavy computation cost. (2), To improve the efficiency, people attempt to compute MHSA within the local/windowed region like CNNs do, for example Swin Transformer [32], Cross-Swin Transformer [33], and TransCNN [29]. (3), Using a CNN-based module (i.e., Token Pyramid Module), the TopFormer model proposed in [34] takes Tokens from various scales as input to produce scale-aware semantic features. 3M-CDNet-V2 provides a different look: it directly replaces the spatial deformable convolutions with MHSA in the final layers of a CNN backbone, and our aim is to obtain rich semantics and large receptive field.

In this article, we make the following contributions:

(1) Inspired by the recent success of the light-weight CD network: 3M-CDNet, we made a series of modifications on it, which brings a significant performance boost.

(2) We replace the spatial deformable convolutions with MHSA in the final layers of 3M-CDNet, so as to learn long-range dependencies in the bitemporal images.
A new transfer learning strategy was applied to facilitate the application of 3M-CDNet-V2 on small datasets.

The rest of this article is organized as follows. Section II describes the related work on 3M-CDNet and gives the details of our proposed architecture. The comparative experimental results and ablation studies are reported detailedly in section III. The discussion is given in section IV and conclusion is drawn in section V.

II. METHODOLOGY

A. A BRIEF INTRODUCTION OF 3M-CDNET

As shown in Fig. 1, 3M-CDNet mainly consists of four components: Input Layer, Layer 1, Layer 2, and Classifier. The former three are used to extract deep features from the input image $I^{(1,2)} \in \mathbb{R}^{6 \times H \times W}$ which is a concatenation of the pre-change image and post-change image along the dimension of channels. Therein, as shown in Figure 2, Layer 1 and 2 are “modulated deformable convolution” (MDConv)-based backbone networks [35]. The latter one which is a pixel-wise classifier is used to classify the extracted features into two classes in a change probability map [3]. Input of the model is a six-band fused RGB image, and the output is a binary change map $CM \in \mathbb{R}^{1 \times H \times W}$, where pixels are either unchanged or changed. 3M-CDNet is a lightweight CD network for remote sensing images, and it only involves about 3.12M trainable parameters.

Using 3M-CDNet, change detection can be implemented based on high-resolution feature maps (at least 1/4 of the original image size). Reducing the width and depth of the backbone network also decreases the number of model parameters, thus creating high-resolution and lightweight deep features. At the same time, however, the receptive field of these deep features is limited due to the limitation of the backbone network’s depth, and the completeness of the contextual semantics in deep features cannot be kept [36], leading to weak feature representation. Fortunately, MDConv, which is a powerful mechanism to attend to flexible spatial locations conditioned on input data [35], is designed to capture deformable context from objects with various shapes and scales, thus providing a feasible solution to alleviate these problems. In other words, learning a deformable receptive field for the convolution filters has been proved effective in selectively attending to more informative regions in the feature maps [37].

B. MODIFICATIONS TO 3M-CDNET

Comparing Fig. 1 & 2 to Fig. 3, we have made nine important modifications to the original version of 3M-CDNet, they are:

![Diagram](image-url)
1) CHANGE THE EARLY-FUSION FRAMEWORK AS LATE-FUSION
In Fig. 1, 3M-CDNet adopts the early fusion strategy to treat change detection issues as exclusively semantic segmentation problems. This strategy [1], however, has been proven to be sensitive to noisy conditions such as geometric distortion and different viewing angles, and often leads to dropped accuracy [1]. Inasmuch, as displayed in Fig. 3a, we modified the input layer to receive bitemporal images separately instead of two concatenated images as done in Fig. 1. To be specific, we changed the original early-fusion framework as late-fusion by constructing a Siamese network with shared weights, in which the resulting features are extracted from the fused results (i.e., subtraction) of two “independent” branches.

2) CHANGE “CONV” AS “MBCONV” IN THE INPUT LAYER
We used a customized CNN module to extract features from image patches. As illustrated in Fig. 3b, from 3M-CDNet to 3M-CDNet-V2, two $3 \times 3$ convolutions in the Input Layer were replaced by “MBConv” which denotes the mobile inverted bottleneck convolution [36]. MBConv is widely used in EfficientNet [36], MobileNet [38], and many other lightweight backbone networks. Comparing to its original version, here we added a residual connection, and removed the SE (Squeeze-and-Excitation)-attention layer for smaller parameters and higher accuracy. An important aspect of MBConv is that it creates an inverted bottleneck, i.e., its hidden dimension is 4 times wider than the input dimension (from $C$ to $4C$). It is reported that wider networks tend to be able to capture more fine-grained features and are easier to train [36].

3) ADD “ASPP” TO THE INPUT LAYER
We also added the ASPP (i.e., Atrous Spatial Pyramid Pooling) module [39] to the Input Layer. ASPP is introduced for the purpose of obtaining several multi-scale receptive fields, which may significantly improve the networks’ ability to handle both large and small changes [39]. The introduction of ASPP has also deepened the Input Layer without reducing the size of the feature maps. The benefit of deeper networks is that they can capture richer, more complex [36], and highly discriminative features, and generalize well on new datasets. By dense skip connections, the output features of the Input Layer ($X_0$) will then be fed into subsequent modules to determine the change.

4) THE WIDE USE OF SKIP CONNECTION
Much evidence demonstrates that both high-level semantics and low-level detail information are important in change detection [1], so, as shown in Fig. 3a, the skip connections are widely used in our architecture to fuse the feature representations at different levels. For example, the input of Layer 2 is a fused feature representation obtained by concatenating the high-level $X_1$ and low-level $X_0$ along the channel axis, and the input of Classifier 1 is a representation obtained by fusion.

FIGURE 2. Architecture of the MDConv bottleneck residual blocks (Modified after [11]). Therein, convolution layers are denoted by “number of kernels of each filter, kernel size, number of filters”, e.g., “128, Conv1 $\times 1$, 64”. Conv and MDConv indicate the 2-D convolution layer and modulated deformable convolution layer, respectively.
or concatenation of $X_0$, $X_1$, and $X_2$. In the original version of 3M-CDNet, there is only one skip connection sending the $X_1$-$X_2$ fused representation to Classifier, and the authors claimed that more connections seem helpless in improving the model performance [11]. However, in our case, the opposite strategy, which can take full advantage of the improved feature representation $X_0$ via dense skip connections, was proved to work better.

5) THE WIDE USE OF DEEP SUPERVISION
Multistage prediction and deep supervision can improve the learning ability of the feature extractors and thus help derive

---

**FIGURE 3.** Pipeline of the 3M-CDNet-V2 architecture. Note: red ellipses represent important modifications to 3M-CDNet. The legends in this figure are consistent with those in Fig. 1 & 2, if not specified.
highly discriminative information [5]. As shown in Fig. 3a, 3M-CDNet-V2 introduces deep supervision between each “Layer” pair, and four deep supervision branches are used in 3M-CDNet-V2. The implementation of deep supervision is as follows: taking the Input Layer as example, first calculate \( X_{\text{difference}} = |X_0 - X_1| \), where 1 and 2 represent time \( t_1 \) and \( t_2 \), respectively; and then, use a convolution layer to reduce the dimension of \( X_{\text{difference}} \) to 1 (i.e., the number of classes); subsequently, upsample \( X_{\text{difference}} \) to the original image size; finally, calculate the loss between the predicted confidence map \( X_{\text{difference}} \) and ground truth during the training process. In this way, the intermediate layers are effectively trained and weights of them can be finely updated [16], thus alleviating the presence of vanishing gradient and improving the model performance. Deep supervision strategy facilitates the network convergence during the training phase, whereas it brings about more computation and memory cost than single-head prediction [11], so the original 3M-CDNet did not apply it. However, we will prove that deep supervision is crucial in 3M-CDNet-V2.

6) THE INTRODUCTION OF MHSA IN LAYER 2
This idea is borrowed from [31]. Authors in [31] have proven that by just replacing the spatial convolutions with global multi-head self-attention (MHSA) in the final three bottleneck blocks of a ResNet and no other changes, their approach improves upon the baselines significantly on object detection and instance segmentation while also reducing the parameters, with minimal overhead in latency. In a similar fashion, as shown in Fig. 3c, we replaced the spatial convolutions with MHSA in the final three bottleneck blocks of Layer 2 which have the minimum feature-map size of H/8 and W/8. By leveraging the MHSA, Layer 2’s MDConvBottleneck has larger receptive fields and is notable for its use of attention to model long-range dependencies in the data. In the original 3M-CDNet, the authors attempt to use MDConv, rather than down-sampling operations, to enlarge the receptive field of deep features. But it still has a difficulty in producing high-quality classifications because it lacks the element relation modeling mechanism which is the key for the success of image-processing tasks [37]. The network architecture of MHSA used in Fig. 3c is shown in Fig. 4. Note that: in order to make the self-attention position-aware, a relative-distance-aware position-encoding [40] module which is better suited for vision tasks [41] are incorporated in MHSA.

8) INTRODUCTION OF LOW-LEVEL FEATURES TO THE CLASSIFIER
It is believed that the high-level feature representations in CNN are accurate in semantics but coarse in location, and
The legends in this figure are consistent with those in Fig. 1, 2 & 3, if not specified.

9) REMOVE "SIGMOID + BINARIZATION" OPERATION FROM CLASSIFIER

For the sigmoid/softmax function which maps any real values to the range [0, 1] has already been embedded in our loss functions (see Section 2.3), we removed the "sigmoid + Binarization" operation shown in Fig. 1 from the Classifier 2 in order to avoid the vanishing gradients caused by repetitive normalization [5]. On the other hand, we reuse the sigmoid/softmax function when calculating the evaluation metrics in order to scale all the predicting data between 0 and 1. However, as the sigmoid/softmax module no longer accepts the gradient-descent update, we cannot use the commonly used threshold 0.50 [5] to perform the binary classification, and the threshold value must be manually adjusted.

### C. THE LOSS FUNCTION

In the practical applications of change detection, the number of unchanged objects is often much more than that of changed ones. At the suggestion of [15], one way to alleviate the impact of sample imbalance is to introduce a hybrid and weighted loss function [5]. In this article, we consider using the combination of the Focal loss [17] and Dice loss [42]. The hybrid loss function \( L \) can be formulated as follows:

\[
L = L_{\text{dice}} + L_{\text{focal}}
\]

Therein, the Dice loss \( L_{\text{dice}} \) attaches similar importance to false negatives and false positives, and it is effectively immune to the sample-imbalance issue [42].

\[
L_{\text{dice}} = \frac{1}{U - I + \epsilon} \sum_{i} (y_{1} - y_{2})
\]

\[
I = \sum_{i} t_{i} y_{i}
\]

\[
U = \sum_{i} (t_{i}^2 + y_{i}^2)
\]

where the smoothing factor \( \epsilon \) is an extremely small number (here \( \epsilon \) is set as 1e-7); \( 0 \leq y_{i} \leq 1 \) is the predicted value processed using the sigmoid/softmax activation function; and \( t_{i} \) is the ground-truth value, being either 0 or 1.

The \( L_{\text{focal}} \) loss is defined as:

\[
L_{\text{focal}} = -\alpha_{t} (1 - p_{t}) \gamma \log(p_{t})
\]

where \( p_{t} \) is the probability that the model predicts for the ground truth objects. Focal loss adds \((1 - p_{t}) \gamma \) as a modulating term to the standard cross entropy loss so that it can focus learning on hard examples and prevents the large number of easy negatives from overwhelming the detector during training. When \( p_{t} > 0.5 \), setting \( \gamma > 0 \) reduces the relative loss for well-classified samples, thus putting more focus on those hard and misclassified samples. What’s more, \( L_{\text{focal}} \) gives high weights \((\alpha_{t})\) to the rare class and small weights
to the dominating or common class. Here we empirically set \( \gamma = 2 \), and \( \alpha_t = 0.25 \sim 0.45 \).

### III. EXPERIMENTS
#### A. RELEVANT DATA SETS
To evaluate our medium-weight CD model, a series of comparative experiments and ablation experiments were conducted on CDD (Change Detection Dataset) and PIESAT-CD, and CD_Data_GZ.

CD is a classic dataset for testing the performance of CD models. It contains 11 pairs of RGB images taken in different seasons obtained by Google Earth [5], and the spatial resolution of them is 3 cm/px to 100 cm/px. After cropping from the original high-resolution images, this dataset is split into 10000 image pairs for training, 3000 pairs for validation, and 3000 pairs for testing, respectively, and the size of each image pair is 256 \( \times \) 256 pixels.

PIESAT-CD is an open-source dataset released in 29-10-2021 at the 4th Chinese Conference on Pattern Recognition and Computer Vision. The dataset is available at https://captain-whu.github.io/PRCV2021_RS/dataset.html. Based on the Gaofen-2 RS images covering the main busy cities of China, PIESAT-CD provides 6000 bitemporal image pairs of size 512 \( \times \) 512 \( \times \) 3 in .png format for training, 2000 pairs for validation, and 2000 for testing. However, due to the introduction of MHSA in 3M-CDNet-V2, a 512 times 3 in .jpg format. And then, we removed those non-changed image pairs. As a result, 19030 non-overlapping image pairs were cropped for training, 7000 pairs for validation, and 7200 for testing. The spatial resolution of them is approximately 0.8\( \sim \)1.0 m/px. PIESAT-CD only detects changes in the buildings, while other changed objects are omitted from the analysis. The ground truth labels were manually annotated by RS image interpretation experts who are employed by the PIESAT International Information Technology Limited. Each sample in the dataset was annotated by one expert and then double checked by another to produce a high-quality CD dataset. Even so, we still found that there are quite a few inaccurately labeled, unregistered or dual-view image pairs in the training dataset, which may hurt the predicting accuracies.

The CD_Data_GZ dataset is an open-source dataset released by [13]. It includes Google image pairs covering the outskirts of Guangzhou, China from 2006 to 2019, as well as detection tags of changed buildings. CD_Data_GZ provides 20 RGB image pairs of size 1006\( \times \)1168 to 4936\( \times \)5224, and their spatial resolution is 0.55m/pixel. In this research, they were cropped into 3860 non-overlapping tiles of size 256 \( \times \) 256 \( \times \) 3 in .jpg format. And then, we removed those non-changed image pairs to avoid the “change versus non-change” imbalance phenomenon. Finally, 1110 image pairs with positive and negative instances were used for training, and the remaining 206 image pairs used for testing.

These three datasets were collected from several platforms, sensors, and areas/cities, focusing on different change type scenarios, so it is better to use them to test the applicability, robustness, and generalization ability of 3M-CDNet-V2. In addition, (1): Google Earth images and Gaofen-2 images are the preferred datasets in engineering applications because of their low cost and wide availability in the market. (2): There are many erroneously annotated labels in PIESAT-CD and CD_Data_GZ, but this scenario is particularly common in engineering applications. From the application-oriented perspective, these are the reason why CDD, PIESAT-CD and CD_Data_GZ were selected to train and verify our model.

#### B. EVALUATION METRICS
The Classifier 2 module shown in Fig. 3d is a single-class object classifier, and it produces only one change map for each input image pair (the number of classes is 1). After the change maps are normalized between 0 and 1 using the sigmoid/softmax function, we need a binarization threshold defining whether a certain pixel in it is changed or not. With the purpose of maximizing the Precision and Recall at the same time and keeping a balance between them, this threshold is empirically set as 0.65 or so by trial and error.

By comparing the binarized change map and ground truth, four metrics were used to validate effectiveness and accuracy of the proposed methods, which are: Precision, Recall, F1-Score, and IOU (Intersection over Union). The formulas are given below:

\[
\text{precision} = \frac{TP}{TP + FP}, \quad \text{recall} = \frac{TP}{TP + FN}, \quad F1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \quad \text{IOU} = \frac{TP}{(TP + FP + FN)}
\]

where TP, FP and FN denote the total pixel number of true positive, false positive, and false negative [5], respectively.

#### C. IMPLEMENTATION DETAILS
We implement 3M-CDNet-V2 as well as the relevant comparative models using the Pytorch framework which is open source. We conducted these experiments on a single NVIDIA Tesla V100 with 32 GB of GPU memory and train for 100 epochs to make the models converge. The relevant models were trained from scratch, and the batch size is set to 10. The AdamW optimizer is applied to train our models. The initial learning rate is set to 1e-3, and we decay the learning rate of each parameter group by gamma (= 0.770) every ten epochs. Validation is performed after each training epoch, and the best model on the validation set is saved in .pth format and used for evaluation on the test set. We also applied online data augmentation to the input image patches, including flip, normalization, rotate, gaussian blur, histogram matching, shuffle, etc.
In addition, as the CD_Data_GZ dataset is too small to fit the experimental models, so we introduced the transfer learning technology here. To be specific, before trained on CD_Data_GZ, we initialize 3M-CDNet, 3M-CDNet-V2, and several other compared models using the weights pretrained on the PIESAT-CD dataset (shown in Table 1), instead of randomly initializing them. In theory, this operation could help our neural networks reaching better and robust classification performance [5] on CD_Data_GZ, and reducing the risk of over-fitting.

D. COMPARISON AND ANALYSIS
1) THE PERFORMANCE OF THE PROPOSED MODEL
In this section, we will compare the model performance of our 3M-CDNet-V2 with several existing SOTA methods as follows:

(1): FC-Siam-Diff [10] is a feature-difference method, which extracts multi-level features of bi-temporal images from a siamese UNet architecture, and their difference is used to detect changes.

(2): STANet [9] is a siamese-based spatial-temporal attention network for CD. Note that STANet is fine-tuned on the ImageNet-pre-trained ResNet-18 model.

(3): IFN [16] can fuse multi-scaled deep representations of bi-temporal images with image difference representations by means of attention modules for change map reconstruction. It uses pre-trained VGG16 as the backbone network.

(4): ChangeSTAR [43] is a single-temporal supervised change detection algorithm. To ensure the fairness, in this paper we still use bitemporal image pairs as its input. Note that ChangeSTAR uses pre-trained deep network (ResNet-101) as the encoder. During the training process, we set the number of iterations as 5600, and the batch size is set as 4 for input images of size 512 × 512, as 16 for images of size 256 × 256, and the SGD optimizer with momentum=0.9 and weight_decay=0.0001 is used.

(5): SNUNet-CD [17] applies a densely connected Siamese network, namely the Siamese NestedUNet network [5], for change detection.

(6): SIUNet3+CD [5] is a siamese UNet3+-based CD network architecture. Note that it uses the GN (Group Normalization) layer [5] to normalize the inputs to a hidden layer, so that the batch size can be set as small as possible in conditions of insufficient computational resources.

(7): BIT [24] is a ResNet-Transformer combined architecture for CD. It applies a transformer encoder-decoder network to enhance the context-information of ResNet features via semantic tokens followed by feature differentiating to obtain the change map. Note that in [24], the number of training epochs is set as 200 to achieve a relatively good performance, whereas for a fair comparison, it is set as 100 in our experiments.

(8): ChangeFormer [25] is a pure Transformer-based siamese network for change detection applications. Specifically, it uses a hierarchical Transformer encoder in a Siamese architecture with a simple MLP decoder.

| Method | Precision | Recall | F1-Score | IOU   |
|--------|-----------|--------|----------|-------|
| FC-Siam-Diff | 0.6290 | 0.6920 | 0.5335 |       |
| STANet | 0.8920 | 0.7723 |       |       |
| IFN | 0.9030 | 0.8237 |       |       |
| ChangeSTAR | 0.9205 | 0.8527 |       |       |
| SNUNet-CD | 0.9350 | 0.9093 |       |       |
| SIUNet3+CD | 0.9069 | 0.8296 |       |       |
| BIT | 0.9331 | 0.8747 |       |       |
| ChangeFormer | 0.8479 | 0.4217 |       |       |

ChangeFormer has a pretrained model which can be downloaded from http://www.dropbox.com/s/undtrlxiz7 bkag5/pretrained_changeformer.pt?dl=0. Note that: to make a fair comparison, here we set the training epochs of ChangeFormer as 100, rather than 200 as recommended by authors in [25].

To enable a fair comparison, whenever possible, we adopt the reported performance from the corresponding publications. Otherwise, if not specified, we re-implement the algorithms based on public codes or using our own Python codes with the parameter settings consistent with the relevant references. The same online data-augmentation strategy as above-mentioned was applied to all the compared methods.

As can be seen from Table 1:
For the CDD dataset, SNUNet-CD achieves the best CD performance. 3M-CDNet-V2 achieves the second-best CD performance in terms of F1, IOU, and Recall metrics compared to other models, including the models with pretrained backbone networks, e.g., STANet, IFN, BIT, ChangeSTAR, ChangeFormer, etc. The BIT model with the pretrained ResNet as its backbone achieves the third best performance. Note that: according to [24], BIT achieves better results on CDD than Table 1 provided. This is because in [24], the BIT model has been trained for 200 epochs; while in Table 1,
FIGURE 6. Visualized result of 3M-CDNet-V2 and the compared methods on the CDD dataset. Red rectangles highlight typical false negatives (the missing portions) or false positives. Note that: due to limited spaces, we could not show the prediction results of all the algorithms listed in Table 1.

the training was stopped after 100 epochs to ensure the fairness. 3M-CDNet gets the worst performance, although its batch size is set as 20 during the training process. Such favorable accuracy is not only significant in quantitative assessment reported in Table 1, but is also visually noticeable from results shown in Fig. 6. To be specific, our proposed 3M-CDNet-V2 model produces less false positives and negatives, and it successfully returns the changed areas with relatively accurate shapes/boundaries and high object internal compactness, although there are some important details still missing from the predicted maps. 3M-CDNet-V2 also has certain capability to trace small changes (e.g., row 6 in Fig. 6) or narrow shaped objects (e.g., row 2, 3, & 5 in Fig. 6). By the way, although SiUNet3+-CD has relatively low F1-score and recall, it is able to identify each small changed object with a clear and separated boundary, showing a slightly better visual performance than ours. However, its computational cost is very expensive – its FLOPs (floating point operations) come up to 216.72G [5] when the bitemporal input-image-size is $2 \times 256 \times 256 \times 3$, while the FLOPs of our V2 model are only 32.76G, thus achieving a good balance between computational efficiency and prediction accuracy.

For PIESAT-CD, 3M-CDNet-V2 still attains the best scores (F1 and IOU) compared with other nine architectures, although there are no pretrained parameters involved in the training process. IFN and 3M-CDNet get the worst performance. Although many of the PIESAT-CD image pairs are not accurately labeled, the precision index achieved by 3M-CDNet-V2 is still above the 85.0% level which is commonly recommended for industrial applications. As shown in Fig. 7, change masks generated by 3M-CDNet-V2 preserve the actual shape of changed objects with complete and separated boundaries, even when there are misregistration errors, clumped masks, and missing objects in the labels. On the other hand, the compared methods including the ones pretrained on the ImageNet dataset, e.g., STANet, ChangeFormer, ChangeSTAR, & BIT, often produce false alarms, fragmentized boundaries, and clumped masks. SiUNet3+-CD and SNUNet-CD, which have very good prediction performance on the CDD dataset, did not present a satisfactory performance on PIESAT-CD, suggesting a limited generalization ability/robustness. Another interesting observation is that although ChangeSTAR has higher prediction accuracy next only to 3M-CDNet-V2, its visual performance is sub-optimal. The case of BIT is similar, indicating it is less robust to pseudo-change and noise.

For CD_Data_GZ, IFN and STANet get the worst performance. 3M-CDNet-V2 achieves the third-best CD performance in terms of precision, recall, F1-score and IOU, next only to ChangeSTAR and SNUNet. However, as illustrated in Fig. 8, the visual performance of ChangeSTAR and SNUNet is quite poor, being vulnerable to pseudo-changes and noise.
FIGURE 7. Visualized result of 3M-CDNet-V2 and the compared methods on the PIESAT-CD dataset. The yellow box indicates false positive annotations in the label, while the red box highlights the detailed information of false negatives or false positives. Note that: due to limited spaces, we could not show the prediction results of all the algorithms listed in Table 1.

Their relatively high recall values do not indicate successful detection of the targets, but lead to many false positives. In addition, there are quite a few missing objects in the ground-truth images, and ChangeSTAR & SNUNet fail to identify them. On the contrary, the visual performance of 3M-CDNet-V2 is improved significantly. As highlighted in red, our 3M-CDNet-V2 captures much finer details compared to other SOTA methods, including the small objects missed in the ground-truth images. Obviously, our proposed method could achieve a good generalization capacity on small datasets by initializing the model parameters using the parameters pretrained on PIESAT-CD, so it is greatly helpful in engineering applications.

The above results demonstrated that 3M-CDNet-V2 has excellent generalization ability and has a strong ability to counter against pseudo-changes and noise.

2) THE EFFICIENCY EVALUATION

Fig. 9a records different models’ parameters versus FLOPs, and 9b records F1 Scores versus FLOPs. There are several useful observations from Fig. 9:

(1): Our V2 model has 18.21 M trainable parameters, and the corresponding FLOPs is 37.63G when the input image size is 256 times 256. The original 3M-CDNet model has 3.29M parameters and 23.69G FLOPs. The increased number of parameters mainly come from the introduction of MHSA, see Table 2 below.

(2): ChangeSTAR, IFN, ChangeFormer, and SiUNet3+-CD have the maximum trainable parameters, and therein, IFN and SiUNet3+-CD have the largest number of FLOPs: 82.26 G and 216.72 G. Although some of them could provide an acceptable CD performance, the expensive computational cost limits their real-world applications.

(3): FC-Siam-Diff, 3M-CDNet, and SNUNet-CD have the least number of trainable parameters, but they offer significantly lower CD accuracies in our experiments. Even worse, when the input image size is 256 × 256, the FLOPs of 3M-CDNet and SNUNet-CD come up to 23.69G and 54.77G, respectively. Thus, these cannot provide a good balance between computational complexity and performance.

(4): BIT, STANet, and 3M-CDNet-V2 have medium parameters, and their FLOPs are 10.59G, 12.98G, and 37.63G, respectively. Referring to Table 1, we conclude that BIT achieves an optimal trade-off between efficiency and accuracy, and the proposed 3M-CDNet-V2 achieves the second- or third-best trade off. Obviously, it is the MHSA module that gives 3M-CDNet-V2 a slightly better CD result while moderately increasing the computational complexity.

We also checked the training speed (time consuming) of different models, which is in general consistent with...
FIGURE 8. visualized result of 3M-CDNet-V2 and the compared methods on the CD_Datagz dataset. The yellow box indicates false annotations in the label, while the red box highlights the detailed information for comparing purpose. Note that: due to limited spaces, we could not show the prediction results of all the algorithms listed in Table 1.

FLOPs. For example, if setting batch size = 2, it costs 49s for FC-Siam-diff to run a forward iteration 200 times during the training phase; for ChangeFormer (with pretrained model) it costs 47s; for SNUNet-CD it costs 1 min and 30s; for our 3M-CDNet-V2 it costs 2 mins and 04s; for IFN it costs 2 mins and 12s; and for SiUNet3+-CD it costs 4 mins and 29s. Although our model training is relatively time-consuming, its inference efficiency is satisfactory. If setting batch size = 20 and the input image size = 256 × 256, it takes only 1.0s for our model to make one prediction.

3) ABLATION STUDIES – TESTING ON THE PIESAT-CD DATASET
In this section, we study the effect of removing different components to provide insights into what makes the proposed V2 model successful. We make the quantitative comparisons on the PIESAT-CD dataset. To be specific, we conduct experiments in Table 2 to ablate each component of 3M-CDNet-V2 step by step. To ensure a fair comparison, the parameter settings of different models in Table 2 are devised to be the same during training, for example, the batch size of them is fixed as 10.

We can make the following observations from Table 2:
1. A steady upward trend in accuracy metrics and computational efficiency is visible from “O” to “O+SUP+LF+MBConv+ASPP+SC+MHSA+τ”.
2. There is a significant accuracy boost when adding deep supervision modules to 3M-CDNet (“O”). To be specific, the F1 score of “O+SUP” is 23.64% higher than the baseline model “O”.
3. Another significant boost in model prediction accuracy occurs when using the late fusion approach. To be specific, the F1 score of “O+SUP+LF” is 10.88% higher than “O+SUP”, and 34.52% higher than “O”.
4. Only a slight upward trend in F1 score and IOU can be observed when the MBConv and ASPP modules were incorporated into the baseline model.
5. The incorporation of “SC”, “MHSA”, and “τ” could reinforce the performance of the baseline model, and they achieve substantial accuracy gains. For example, the F1 score of “O+SUP+LF+. . .+SC” is 1.23% higher than “O+SUP+LF”, the F1 score of “O+SUP+. . .+MHSA” is 0.93% higher than “O+SUP+LF+. . .+SC” and 2.16% higher than “O+SUP+LF”, and the F1 score of 3M-CDNet-V2 (“O+SUP+LF+. . .+τ”) is 1.22% higher than “O+SUP+LF+. . .+MHSA” and 3.38% higher than “O+SUP+LF”.

Accordingly, in Fig. 10, the visual quality of the baseline models “O” and “O+SUP” is quite poor, while the incorporation of “LF” helps rebuild the deep features and keep the boundary of the change map continuous. The visual performance of “O+SUP+. . .+MBConv” is slightly better than “O+SUP+. . .+ASPP”. The introduction of “SC” further enriched the representation of deep feature information, but there are still quite a few false alarms, especially in Row 3, 4, and 5. The introduction of “MHSA” can
learn the features of the changing objects better, with reduced incorrect detection while partly avoiding the phenomenon of “clumping” in the detection areas. 3M-CDNet-V2 gives the optimal visual performance: complete boundaries, few clumped masks, and fewer false detections. This is because the introduction of “τ” guides the network to learn more detailed difference information between the bi-temporal images.

From the above, we can conclude that our modifications push the baseline models up to new state-of-the-art records, and trading the improved performance with a certain increase in computation costs is therefore worthwhile.

IV. DISCUSSION

In our opinion, there are eight meaningful aspects around the proposed network architecture worth to discuss, which are listed in the following paragraphs.

First, among all the modifications we made to 3M-CDNet, the late fusion framework and deep supervision have been proven to contribute most to the performance gain, followed by “SC” and MHSA. However, authors in [11] argue that: without deep supervision and the connections skipped from the Input Layer block, 3M-CDNet still outperforms other SOTA approaches in terms of the comprehensive metrics F1 score and IOU by a significant margin. This conclusion is not supported by the outcome of our study, and as shown in Table 1 and 2, 3M-CDNet has the worst performance among the compared methods. As the source code of 3M-CDNet in [11] is not released to public, we cannot obtain the specific designs and implementation details of this model from the perspective of programming, but one thing is certain: even though 3M-CDNet could achieve upwards to 90% prediction accuracy in CD tasks, further modifications on it are still possible and necessary.

Second, at the center of 3M-CDNet is the high-resolution feature extraction and analysis, which enables the network to reduce the loss of information during the process of down-sampling feature learning. Relative to 3M-CDNet, our model goes deeper: it keeps using high-resolution feature maps in the model training and prediction by incorporating the skip connection between τ and the classifier, which is not only helpful for further improving the CD accuracy metrics, but also has the ability to improve the visual performance of the predicted results, namely promoting the detection of detailed spatial changes and object boundaries.

Third, people may also question that: due to the limited performance gains, why are MBConv and ASPP essential to 3M-CDNet-V2? Actually, before the introduction of “SC”,

![FIGURE 9. (a) Different models' parameters versus FLOPs; (b) F1 Scores versus FLOPs.](image-url)
TABLE 2. The ablation study based on the PIESAT-CD dataset.

| Method                      | Parameters | Precision | Recall | F1-Score | IOU   |
|-----------------------------|------------|-----------|--------|----------|-------|
| O                           | 3.2930M    | 0.4596    | 0.4531 | 0.4563   | 0.2956|
| O+SUP                      | 3.2930M    | 0.7042    | 0.6816 | 0.6927   | 0.5299|
| O+SUP+LF                   | 3.2932M    | 0.8090    | 0.7942 | 0.8015   | 0.6688|
| O+SUP+LF+MBConv            | 3.4259M    | 0.8249    | 0.7999 | 0.8018   | 0.6691|
| O+SUP+LF+MBConv+ASPP       | 3.9837M    | 0.8255    | 0.7871 | 0.8059   | 0.6749|
| O+SUP+LF+MBConv+ASPP+SC    | 4.3715M    | 0.8238    | 0.8040 | 0.8138   | 0.6860|
| O+SUP+LF+MBConv+ASPP+SC+MHSA | 17.9147M  | 0.8469    | 0.8005 | 0.8231   | 0.6993|
| O+SUP+LF+MBConv+ASPP+SC+MHSA+τ | 18.2096M  | 0.8522    | 0.8191 | 0.8353   | 0.7172|

Best results are highlighted in bold font. Herein, O represents the original 3M-CDNet, SUP: deep supervision, LF: late fusion, SC: skip connections, and τ represents the introduction of low-level feature τ by skip connection.

FIGURE 10. Visualized results of the ablation experiment on the PIESAT-CD dataset. The red box highlights the detailed information for comparing purpose. The yellow box indicates false annotations in the label.

it is Layer I and Layer II that contribute most to the final results, and thus the accuracy gains obtained by MBConv/ASPP are not impressive. However, after the introduction of “SC”, the features extracted by MBConv and ASPP become part of the input of Layer I, Layer II, and Classifier 1, and then we get substantial accuracy gains. Authors in [11] found that the connections skipped from the Input Layer block have nearly no improvement in the CD performance, but when MBConv + ASPP were added, the skip connections begun to show its effects. This explains why MBConv and ASPP are essential to 3M-CDNet-V2.

Fourth, MDConv is of an efficient mechanism to enhance the geometric transformation modeling ability [11] of deep feature representations, and it can also increase the receptive fields of them. However, as pointed out by [44], MDConv lacks the element relation modeling mechanism, which is the key for the success of a CD network architecture. Thus, in 3M-CDNet-V2, MHSA was incorporated into the backbone network immediately after the MDConvBottleneck module. This strategy brings two main benefits: 1. further enlarging the receptive field of deep features; and 2. modeling long range dependencies. Authors in [31] observed a significant increase in prediction accuracy by just replacing the spatial convolutions with MHSA in the final three bottleneck blocks of a ResNet, but we did not – the accuracy gains obtained by MHSA in our model is moderate. In fact, most investigations in the literature [7], [26], [29] show that the Transformer-based architectures could achieve comparable or slightly better results than their CNN counterparts, so the performance of MHSA in our experiment meets the expectations, although its application involves large computational costs. It can be seen from Fig. 10 that the introduction of
MHSA greatly improves the visual quality of CD maps, so it is not optional but essential for achieving high accuracy. In addition, at the suggestion of [7], long training epochs is usually necessary for the MHSA weights to be learned to focus on the changing objects of interest. Our experiments also demonstrated that when increasing the training epochs from 100 to 200, 3M-CDNet-V2 achieves considerable gains (>2.0%) in terms of IOU, F1 score, and precision over the baseline model on the PIESAT-CD dataset. This reminds us that further training may be helpful in further improving the model performance in identifying changes, so the potential of our model might be underestimated.

Fifth, the method proposed in this research is very effective for the dynamic monitoring of changing objects at different scales, but it can be further optimized in the following aspects. (1), as shown in Fig. 3, we use the subtraction operation to fuse the bitemporal feature representations, but in many cases, direct concatenation might be a slightly better choice for feature fusion [45]. (2), it has been demonstrated that our model has good generalization ability and transferability from PIESAT-CD to CD_Data_GZ, but did not generalize well from CDD to CD_Data_GZ. CD_Data_GZ differs from CDD in terms of sensor type and definitions of change. For the purpose of engineering application, it is worthwhile to create a general pretrained model using sufficient training samples [5] covering different sensor types, changing objects, and imaging conditions. (3), one ongoing challenge in the use of 3M-CDNet-V2 for CD is how to automatically set the threshold for binary “change versus non-change” classification. If the threshold value is large, precision is higher and recall is lower, and vice versa. In this article, the optimal threshold values (i.e., 0.65 or so) were determined through a process of trial and error. In the future, we plan to modify the Classifier 2 module in Fig. 3d as a two-class object classifier, so that the threshold values can be determined automatically by the “argmax” function in Python. (4), MDCConv is of an efficient mechanism to attend to sparse spatial locations, and MHSA has powerful element-relation-modeling capacity, so the combination of them in 3M-CDNet-V2 yielded excellent CD results. However, Xia et al. [37] believed that it is possible and necessary to do both in a singular architecture, and a novel deformable self-attention transformer (DAT) was put forward. DAT provides a flexible scheme enabling the self-attention module to selectively focus on important regions and capture more informative deep features. Reference [44] did similar work in the field of object detection. For future research, we plan to simplify the 3M-CDNet-V2 architecture by directly replacing the MDCConv-MHSA combination with DAT, so as to make the model more lightweight, more robust, easier to use, and cheaper.

Sixth, as the core part of Transformers, MHSA is believed to be not suitable for small sets of data [26], while we note that the introduction of transfer learning is necessary. However, unlike the classic transfer strategy, it is better to initialize all the parameters of 3M-CDNet-V2 using the weights pretrained on big datasets and then fine-tune them, rather than freezing and only initializing the encoder part. This is an important finding of our study, and it allows for a MHSA-based model to be trained on very small datasets like CD_data_GZ.

Seventh, as the F1 score vs. efficiency of our model is not better than some existing methods such as BIT, so seemingly, the only contribution of our model is to detect narrow and little changes. But in a strict sense, we may need to collect more evidence or conduct more experiments to support this claim against other existing methods. Actually, we accept that our 3M-CDNet-V2 may not be the optimal solution for change detection, but it does provide competitive or slightly better accuracy metrics or shows better visual quality (than the compared methods including BIT) on three known datasets. This fact confirms the benefits of the modifications we made to 3M-CDNet, and in particular, it improves our understanding of architecture design spaces for deep supervision, late fusion, MHSA, and τ in RS change detection. In other words, with improved detection sensitivity, our model is a step forward in designing an effective end-to-end change detection technique for bitemporal high-resolution image analysis, and this is the most important contribution of the article. We also believe that, due to the incorporation of MHSA and its power at modeling global contexts [7], the detection sensitivity and accuracy of our model on the new dataset or in engineering applications could be guaranteed. As for the computational efficiency, we accept that our model is inefficient relative to some of the compared methods, but this is not a major concern in practical applications because our study has demonstrated that 3M-CDNet-V2 has superior transferability for downstream tasks under large-scale pre-training. Given the above, the current study makes meaningful contributions to the knowledge on RS change detection.

By the way, although our 3M-CDNet-V2 have achieved satisfactory performance, it requires an enormous amount of training data, which can cost considerable expense and time. To ease the effort of acquiring high-quality annotation data, in recent years, many studies have focused on unsupervised or self-supervised CD algorithms, for example the style transformation-based spatial-spectral feature learning for unsupervised CD [46], as well as the task-related self-supervised learning for CD, i.e., the discriminative adversarial deep neural networks (DADNN) proposed in [47]. The performance of these algorithms is even comparable to that of the state-of-the-art supervised or semi-supervised methods. In our case, unsupervised 3M-CDNet-V2 might not work in theory because MHSA is data-hungry, but it is feasible to equip 3M-CDNet-V2 or its variants with a self-supervised preprocessing module such as DADNN. In a follow-up study, we are planning to train a similar architecture on different CD datasets to facilitate its engineering application.

V. Conclusion
To satisfy the needs of rapid mapping of change information in high resolution RS imagery, an efficient medium-weight network architecture was proposed for change detection.
This architecture was developed as an updated version of 3M-CDNet by combining CNN and MHSA into a singular network, termed 3M-CDNet-V2. Using three engineering CD datasets (CDD, PIESAT-CD and CD-data_GZ) as benchmark, a series of comparative experiments and ablation experiments were conducted, and our findings showed that: (1) relative to other state-of-the-art algorithms, the proposed algorithm obtained competitive or slightly better CD performance on both visual comparison and quantitative metrics evaluation in terms of F1-Score and IOU. (2) By applying a novel transfer learning strategy, 3M-CDNet-V2 could perform well on the small dataset. (3) 3M-CDNet-V2 maintains a suboptimal trade-off between computational complexity/time and performance. (4) The late fusion framework and deep supervision contribute most to the performance gain of 3M-CDNet-V2, followed by the skip connection and MHSA. In particular, the introduction of MHSA could reinforce the performance of 3M-CDNet while moderately increasing the computational cost. (5) The introduction of the skip connection linking r to the classifier gives 3M-CDNet-V2 a better visual performance, so that it can capture detailed spatial changes such as small changes, narrow shaped objects, and accurate boundaries.

ACKNOWLEDGMENT
Helpful review comments and constructive criticism on this manuscript were provided by several anonymous reviewers, and the authors would like to thank them here. They also thank the PIESAT International Information Technology Ltd., for providing the PIESAT-CD dataset.

REFERENCES
[1] H. Jiang, M. Peng, Y. Zhong, H. Xie, Z. Hao, J. Lin, X. Ma, and X. Hu, “A survey on deep learning-based change detection from high-resolution remote sensing images,” Remote Sens., vol. 14, no. 7, pp. 1–31, 2022.
[2] S. Camalan, K. Cui, V. P. Pauca, S. Alqahtani, M. Siliman, R. Chan, R. J. Plummer, E. N. Dethier, L. E. Fernandez, and D. A. Lutz, “Change detection of Amazonian alluvial gold mining using deep learning and sentinel-2 imagery,” Remote Sens., vol. 14, no. 7, pp. 1–22, Apr. 2022.
[3] W. Wiratama, J. Lee, S. E. Park, and D. Sim, “Dual-dense convolution network for change detection of high-resolution panchromatic imagery,” Appl. Sci., vol. 8, no. 10, pp. 1–13, Oct. 2018.
[4] M. Wang, H. Zhang, W. Sun, S. Li, F. Wang, and G. Yang, “A coarse-to-fine deep learning based land use change detection method for high-resolution remote sensing images,” Remote Sens., vol. 12, no. 12, pp. 1–33, Jun. 2020.
[5] B. Zhao, P. Tang, X. Luo, L. Li, and S. Bai, “SINet3+CD: A full-scale connected Siamese network for change detection of VHR images,” Eur. J. Remote Sens., vol. 55, no. 1, pp. 232–250, Mar. 2022.
[6] X. Huang, D. Wen, J. Li, and R. Qin, “Multi-level monitoring of subtle urban changes for the megacities of China using high-resolution multi-view satellite imagery,” Remote Sens. Environment, vol. 196, pp. 56–75, Jul. 2017.
[7] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, “CVT: Introducing convolutions to vision transformers,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 2–22.
[8] D. Peng, Y. Zhang, and H. Guan, “End-to-end change detection for high resolution satellite images using improved UNet++,” Remote Sens., vol. 11, no. 11, pp. 1–23, Jun. 2019.
[9] C. Hao 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, pp. 1–23, May 2020.

[10] R. C. Daudt, B. L. Saux, and A. Bouchl, “Fully convolutional Siamese networks for change detection,” in Proc. 25th IEEE Int. Conf. Image Process. (ICIP), Oct. 2018, pp. 4063–4067.
[11] K. Song, F. Cui, and J. Jiang, “An efficient lightweight neural network for remote sensing image change detection,” Remote Sens., vol. 13, no. 24, pp. 1–24, Dec. 2021.
[12] F. Jiang, M. Gong, T. Zhan, and X. Fan, “A semisupervised GAN-based multiple change detection framework in multi-spectral images,” IEEE Geosci. Remote Sens. Lett., vol. 17, no. 7, pp. 1223–1227, Jul. 2020.
[13] D. Peng, L. Bresson, Y. Zhang, H. Guan, H. Ding, and X. Huang, “SemiCDNet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 7, pp. 5891–5906, Aug. 2021.
[14] P. Tang, J. Li, F. R. Peng, M. Chen, and X. Li, “PSNet: Change detection with prototype similarity,” Vis. Comput., pp. 1–10, Jun. 2021, doi: 10.1007/s00371-021-02177-4.
[15] W. Zhang and X. Lu, “The spectral-spatial joint learning for change detection in multispectral imagery,” Remote Sens., vol. 11, no. 3, pp. 1–17, Jan. 2019.
[16] C. Zhang, P. Yao, 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 J. Photogramm. Remote Sens., vol. 166, pp. 183–200, Aug. 2020.
[17] S. Fang, K. Li, J. Shao, and Z. Li, “SNUNet-CD: A densely connected Siamese network for change detection of VHR images,” IEEE Geosci. Remote Sens. Lett., vol. 19, pp. 1–5, 2022.
[18] 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 J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 14, pp. 1194–1206, 2021.
[19] X. Yang, L. Hu, Y. M. Zhang, and Y. Q. Li, “MRA-SNet: Siamese networks of multiscale residual and attention for change detection in high-resolution remote sensing images,” Remote Sens., vol. 13, no. 22, pp. 1–21, Nov. 2021.
[20] Z. Cao, M. Wu, R. Yan, F. Zhang, and X. Wan, “Detection of small changed regions in remote sensing imagery using convolutional neural network,” IOP Conf. Ser. Earth Environ. Sci., vol. 502, no. 1, pp. 1–11, May 2020.
[21] C. Wu, B. Du, and L. Zhang, “Slow feature analysis for change detection in multispectral imagery,” IEEE Trans. Geosci. Remote Sens., vol. 52, no. 5, pp. 2858–2874, Jul. 2014.
[22] F. Thonfeld, H. Feihlauer, M. Braun, and G. Menz, “Robust change vector analysis (RCVA) for multi-sensor very high resolution optical satellite data,” Int. J. Appl. Earth Observ. Geoinf., vol. 50, pp. 131–140, Aug. 2016.
[23] Z. Zheng, Y. Zhong, S. Tian, A. Ma, and L. Zhang, “ChangeMask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection,” ISPRS J. Photogramm. Remote Sens., vol. 183, pp. 228–239, Jan. 2022.
[24] H. Chen, Z. Qi, and Z. Shi, “Remote sensing image change detection with transformers,” IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–14, 2022.
[25] W. G. C. Bandara and V. M. Patel, “A transformer-based Siamese network for change detection of VHR images,” 2022, arXiv:2201.01393.
[26] X. Chen, C. J. Hsieh, and B. Gong, “When vision transformers outperform ResNets without pre-training or strong data augmentations,” in Proc. Int. Conf. Learn. Reinforcement (ICLR), Barcelona, Spain, 2022, pp. 1–20.
[27] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “AConvNet for the 2020s,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., New Orleans, LA, USA, Jun. 2022, pp. 1–15.
[28] S. d’Ascoli, H. Touvron, M. L. Leavitt, A. S. Morcos, G. Biroli, and L. Van Gool, “Vision transformers with hierarchical attention,” Proc. 38th Int. Conf. Mach. Learn., San Diego, CA, USA, 2021, pp. 2286–2296.
[29] Y. Liu, Y.-H. Wu, G. Sun, L. Zhang, A. Chhatkuli, and L. Van Gool, “Vision transformers with hierarchical attention,” 2021, arXiv:2106.03180.
[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. H. Zhai, L. Car Raqqa, A. Uszkoreit, and N. Houlsby, “An image is worth 16 words: Transformers for visual recognition,” 2021, arXiv:2106.03180.
32] 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 Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Montreal, QC, Canada, Oct. 2021, pp. 10012–10022.

33] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, “CSWIN transformer: A general vision transformer backbone with cross-shaped Windows,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., New Orleans, LA, USA, Apr. 2022, pp. 12124–12134.

34] W. Zhang, Z. Huang, G. Luo, T. Chen, X. Wang, W. Liu, G. Yu, and C. Shen, “TopFormer: Token pyramid transformer for mobile semantic segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., New Orleans, LA, USA, Jan. 2022, pp. 12083–12093.

35] X. Zhu, H. Hu, S. Lin, and J. Dai, “Deformable ConvNets V2: More deformable, better results,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 9300–9308.

36] M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. Int. Conf. Mach. Learn., Long Beach, CA, USA, 2019, pp. 6105–6114.

37] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, “Vision transformer with deformable attention,” 2022, arXiv:2201.00520.

38] A. G. Howard, M. Zhu, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “MobileNets: Efficient convolutional neural networks for mobile vision applications,” 2017, arXiv:1704.04861.

39] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” 2017, arXiv:1706.05587.

40] P. Shaw, J. Uszkoreit, and A. Vaswani, “Self-attention with relative position representations,” 2018, arXiv:1803.02155.

41] I. Bilan and B. Roth, “Position-aware self-attention with relative positional encodings for slot filling,” 2018, arXiv:1807.03052.

42] X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, and J. Li, “Dice loss for data-imbalanced NLP tasks,” 2019, arXiv:1911.02855.

43] Z. Zheng, A. Ma, L. Zhang, and Y. Zhong, “Change is everywhere: Single-temporal supervised object change detection in remote sensing imagery,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Montreal, QC, Canada, Oct. 2021, pp. 15193–15202.

44] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable DETR: Deformable transformers for end-to-end object detection,” 2020, arXiv:2010.04159.

45] Y. Guo, T. Long, W. Jiao, X. Zhang, G. He, W. Wang, Y. Peng, and H. Xiao, “Siamese detail difference and self-inverse network for forest cover change extraction based on Landsat 8 OLI satellite images,” Remote Sens., vol. 14, no. 3, pp. 1–20, Jun. 2022.

46] G. Liu, Y. Yuan, Y. Zhang, Y. Dong, and X. Li, “Style transformation-based spatial–spectral feature learning for unsupervised change detection,” IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–15, 2022.

47] H. Dong, W. Ma, Y. Wu, J. Zhang, and L. Jiao, “Self-supervised representation learning for remote sensing image change detection based on temporal prediction,” Remote Sens., vol. 12, no. 11, pp. 1–38, Jun. 2020.

BO ZHAO was born in Xianyang, China, in July 1985. He received the bachelor’s degree in geographic information system (GIS) from Northwest Agriculture and Forestry University, Xianyang, in 2008, and the master’s degree in geodetection and information technology and the Ph.D. degree in geochemistry from the China University of Geosciences, Beijing, China, in 2011 and 2015, respectively. From 2015 to 2017, he did his postdoctoral research with Chang’an University, Xi’an, China, and Klagenfurt University, Klagenfurt, Austria. He is currently a Senior Research Fellow with the Nanhu Laboratory, Research Center of Big Data Technology. His current research interests include remote sensing, image processing, AI algorithms, and natural geography.

HUI WANG was born in Leping, China, in August 1995. He received the bachelor’s and master’s degrees in surveying and mapping engineering from the China University of Geosciences, Beijing, China, in 2009, and the Ph.D. degree in surveying and mapping from Chang’an University, Xi’an, China, in 2022. He has long been committed to the commercialization of remote sensing analysis technology and the industrialization of remote sensing application technology, and has made in-depth research on the performance analysis of remote sensing software. He is currently the Vice President of PIESAT International Information Technology Ltd. His research interest includes the research and development of a new generation of remote sensing software pixel information expert (PIE).

PANPAN TANG was born in Jiyuan, China, in September 1985. He received the bachelor’s degree in surveying and mapping engineering from the China University of Geosciences, Beijing, China, in 2008, 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, in 2014. From 2014 to 2020, he worked with the Aerospace Information Research Institute, Chinese Academy of Sciences, as an Associate Researcher. He is currently a Senior Research Fellow with the Nanhu Laboratory, Research Center of Big Data Technology. His current research interests include deep learning-based remote sensing (synthetic aperture radar-SAR and optical) image processing, e.g., image segmentation, classification, and change detection.

XIAOYAN LUO was born in Neijiang, China, in July 1995. She received the bachelor’s and master’s degrees in surveying and mapping engineering from the China University of Petroleum, Qingdao, China, in 2018 and 2021, respectively. She is currently a Junior Researcher with the Nanhu Laboratory. Her research interests include deep learning, intelligent interpretation of remote sensing imagery, and remote sensing object detection.

DONGSHENG LIU was born in Cangzhou, China, in November 1984. He received the bachelor’s degree in computer science from Shanxi Agricultural University, Taigu, China, in 2006, the master’s degree in cartography and geographic information system from China Agricultural University, Beijing, China, in 2009, and the Ph.D. degree in surveying and mapping from Chang’an University, Xi’an, China, in 2022. He has long been committed to the commercialization of remote sensing analysis technology and the industrialization of remote sensing application technology, and has made in-depth research on the performance analysis of remote sensing software. He is currently the Vice President of PIESAT International Information Technology Ltd. His research interest includes the research and development of a new generation of remote sensing software pixel information expert (PIE).

BO ZHAO was born in Xianyang, China, in July 1985. He received the bachelor’s degree in geographic information system (GIS) from Northwest Agriculture and Forestry University, Xianyang, in 2008, and the master’s degree in geodetection and information technology and the Ph.D. degree in geochemistry from the China University of Geosciences, Beijing, China, in 2011 and 2015, respectively. From 2015 to 2017, he did his postdoctoral research with Chang’an University, Xi’an, China, and Klagenfurt University, Klagenfurt, Austria. He is currently a Senior Research Fellow with the Nanhu Laboratory, Research Center of Big Data Technology. His current research interests include remote sensing, image processing, AI algorithms, and natural geography.