TransY-Net: Learning Fully Transformer Networks for Change Detection of Remote Sensing Images

Tianyu Yan, Zifu Wan, Pingping Zhang*, Gong Cheng and Huchuan Lu

arXiv:2310.14214v1 [cs.CV] 22 Oct 2023

Abstract—In the remote sensing field, Change Detection (CD) aims to identify and localize the changed regions from dual-phase images over the same places. Recently, it has achieved great progress with the advances of deep learning. However, current methods generally deliver incomplete CD regions and irregular CD boundaries due to the limited representation ability of the extracted visual features. To relieve these issues, in this work we propose a novel Transformer-based learning framework named TransY-Net for remote sensing image CD, which improves the feature extraction from a global view and combines multi-level visual features in a pyramid manner. More specifically, the proposed framework first utilizes the advantages of Transformers in long-range dependency modeling. It can help to learn more discriminative global-level features and obtain complete CD regions. Then, we introduce a novel pyramid structure to aggregate multi-level visual features from Transformers for feature enhancement. The pyramid structure grafted with a Progressive Attention Module (PAM) can improve the feature representation ability with additional inter-dependencies through spatial and channel attentions. Finally, to better train the whole framework, we utilize the deeply-supervised learning with multiple boundary-aware loss functions. Extensive experiments demonstrate that our proposed method achieves a new state-of-the-art performance on four optical and two SAR image CD benchmarks. The source code is released at https://github.com/Drchip61/TransYNet.

Index Terms—Change Detection, Remote Sensing Image, Vision Transformer, Progressive Attention, Deep Learning.

I. INTRODUCTION

Change Detection (CD) plays an important role in the field of remote sensing. It aims to detect the key change regions in dual-phase remote sensing images captured at different times but over the same scene area. In fact, remote sensing image CD has been used in many real-world applications, such as land-use planning, urban expansion management, geological disaster monitoring, ecological environment protection. However, due to the change regions can be any shapes in complex scenarios, there are still many challenges for high-accuracy CD. In addition, remote sensing image CD by handcrafted methods is time-consuming and labor-intensive, thus there is a great need for fully-automatic and highly-efficient CD.

(*Corresponding author: Pingping Zhang.)

TY. Yan, ZF. Wan, PP. Zhang are with School of Artificial Intelligence, Dalian University of Technology, Dalian, 116024, China. (Email: 2981431354@mail.dlut.edu.cn;2537998622@qq.com;zhpp@dlut.edu.cn)

G. Cheng is with School of Automation, Northwestern Polytechnical University, Xi'an, 710072, China. (Email: gcheng@nwpu.edu.cn)

HC. Lu is with School of Information and Communication Engineering, Dalian University of Technology, Dalian, 116024, China. (Email: lhchuan@dlut.edu.cn)

In recent years, deep learning has been widely used in remote sensing image processing due to its powerful feature representation capabilities, and has shown great potential in CD. With Convolutional Neural Networks (CNN) [1] and Transformers [2], [3], many CD methods extract more discriminative features and have demonstrated good performances. However, previous CD methods still have the following shortcomings: 1) With the resolution improvement of remote sensing images, rich semantic information contained in high-resolution images is not fully utilized. As a result, current CD methods are unable to distinguish pseudo changes such as shadow, vegetation and sunshine in sensitive areas. 2) Boundary information in complex remote sensing images is often missing. In previous methods, the extracted changed areas often have regional holes and their boundaries can be very irregular, resulting in a poor visual effect [4]. 3) The temporal information contained in dual-phase remote sensing images is not fully utilized, which is also one of the reasons for the low performance of current CD methods.

To tackle above issues, in this work we propose a novel Transformer-based learning framework named TransY-Net for remote sensing image CD, which improves the feature extraction from a global view and combines multi-level visual features in a pyramid manner. More specifically, the proposed framework has a Y-shape structure whose input is a dual-phase remote sensing image pair. We first utilize the advantages of Transformers in long-range dependency modeling to learn more discriminative global-level features. Then, to highlight the change regions, the summation features and difference features are generated by directly comparing the temporal features of dual-phase remote sensing images. Thus, one can obtain complete CD regions. To improve the boundary perception ability, we further introduce a pyramid structure to aggregate multi-level visual features from Transformers. The pyramid structure grafted with a Progressive Attention Module (PAM) can improve the feature representation ability with additional inter-dependencies through spatial and channel attentions. Finally, to better train the framework, we utilize the deeply-supervised learning with multiple boundary-aware loss functions. Extensive experiments show that our method achieves a new state-of-the-art performance on four optical and two SAR image CD benchmarks.

The main contributions are summarized as follows:

- We propose a Transformer-based learning framework (*i.e.*, TransY-Net) for remote sensing image CD, which can improve the feature extraction from a global view and combine multi-level visual features in a pyramid manner.
- We propose a novel pyramid structure grafted with a

Progressive Attention Module (PAM) to further improve the feature representation ability with additional inter-dependencies through spatial and channel attentions.

- We introduce the deeply-supervised learning with multiple boundary-aware loss functions, to address the irregular boundary problem in CD.

- Extensive experiments on four optical and two SAR image CD benchmarks show that our framework attains better performances than most state-of-the-art methods.

We note that this work is an extension of its previous conference version [5] with some key improvements as follows: 1) We propose a more powerful PAM with joint spatial and channel attentions. 2) We enhance the multi-level visual features with multi-scale pooling. 3) We provide more discussions with other Transformer-based methods. 4) We add more experimental results to verify the effectiveness of the proposed framework and modules.

II. RELATED WORK

A. Change Detection of Remote Sensing Images

Technically, the task of change detection takes dual-phase remote sensing images as inputs, and predicts the change regions of the same places. Before deep learning, direct classification-based methods witness the great progress in CD. For example, Change Vector Analysis (CVA) [6] is powerful in extracting pixel-level features and is widely utilized in CD. With the rapid improvement in image resolution, more details of objects have been recorded in remote sensing images. Therefore, many object-aware methods are proposed to improve the CD performance. For example, Tang *et al.* [7] propose an object-oriented CD method based on the Kolmogorov–Smirnov test. Li *et al.* [8] propose the object-oriented CVA to reduce the number of pseudo detection pixels. With multiple classifiers and multi-scale uncertainty analysis, Tan *et al.* [9] build an object-based approach for complex scene CD. Although above methods can generate CD maps from dual-phase remote sensing images, they generally deliver incomplete CD regions and irregular CD boundaries due to the limited representation ability of the extracted visual features.

With the advances of deep learning, many works improve the CD performance by extracting more discriminative features. For example, Zhang *et al.* [10] utilize a Deep Belief Network (DBN) to extract deep features and represent the change regions by patch differences. Saha *et al.* [11] combine a pre-trained deep CNN and traditional CVA to generate certain change regions. Hou *et al.* [12] take the advantages of deep features and introduce the low rank analysis to improve the CD results. Peng *et al.* [13] utilize saliency detection analysis and pre-trained deep networks to achieve unsupervised CD. Since change regions may appear in any places, Lei *et al.* [14] integrate Stacked Denoising AutoEncoders (SDAE) with multi-scale superpixel segmentation to realize superpixel-based CD. Similarly, Lv *et al.* [15] utilize a Stacked Contractive AutoEncoder (SCAE) to extract temporal change features from image superpixels, then adopt a clustering method to produce accurate CD maps.

Meanwhile, some methods formulate the CD task as a binary image segmentation task. Thus, CD can be finished in a supervised manner. For example, Alcantarilla *et al.* [16] first concatenate dual-phase images as one image with six channels. Then, the six-channel image is fed into a Fully Convolutional Network (FCN) to realize the CD. Similarly, Peng *et al.* [17] combine bi-temporal remote sensing images as one input for CD. Daudt *et al.* [18] utilize Siamese networks to extract features for each remote sensing image, then predict the CD maps with fused features. The experimental results prove the efficiency of Siamese networks. Furthermore, Guo *et al.* [19] use a fully convolutional Siamese network with a contrastive loss to measure the change regions. Zhang *et al.* [20] propose a deeply-supervised image fusion network for CD. There are also some works focused on specific object CD. For example, Liu *et al.* [4] propose a dual-task constrained deep Siamese convolutional network for building CD. Jiang *et al.* [21] propose a pyramid feature-based attention-guided Siamese network for building CD. Lei *et al.* [22] propose a hierarchical paired channel fusion network for street scene CD. The aforementioned methods have shown great success in feature learning for CD. However, these methods have limited global representation capabilities and usually focus on local regions of changed objects. We find that Transformers have strong characteristics in extracting global features. Thus, different from previous works, we take the advantages of Transformers, and propose a new learning framework for more discriminative feature representations.

B. Vision Transformers for Change Detection

Transformers [23] are firstly proposed for time series tasks, such as natural language processing, speech generation. Recently, they have been applied to many computer vision tasks, such as image classification [2], [3], person re-identification [24], [25] and so on. Inspired by the extreme effectiveness, Zhang *et al.* [26] deploy a Swin Transformer structure [3] with a U-Net [27] for remote sensing image CD. Zheng *et al.* [28] design a deep Multi-task Encoder-Transformer-Decoder (METD) architecture for semantic CD. Wang *et al.* [29] incorporate a Siamese Vision Transformer (SViT) into a feature difference framework for CD. To take the advantages of both Transformers and CNNs, Wang *et al.* [30] propose to combine a Transformer and a CNN for remote sensing image CD. Li *et al.* [31] propose an encoding-decoding hybrid framework for CD, which has the advantages of both Transformers and U-Net. Bandara *et al.* [32] unify hierarchically structured Transformer encoders with Multi-Layer Perception (MLP) decoders in a Siamese network to efficiently render multi-scale long-range details for accurate CD. Chen *et al.* [33] propose a Bitemporal Image Transformer (BIT) to efficiently and effectively model contexts within the spatial-temporal domain for CD. Ke *et al.* [34] propose a hybrid Transformer with token aggregation for remote sensing image CD. Song *et al.* [35] combine the multi-scale Swin Transformer and a deeply-supervised network for CD. All these methods have shown that Transformers can model the inter-patch relations for strong feature representations.
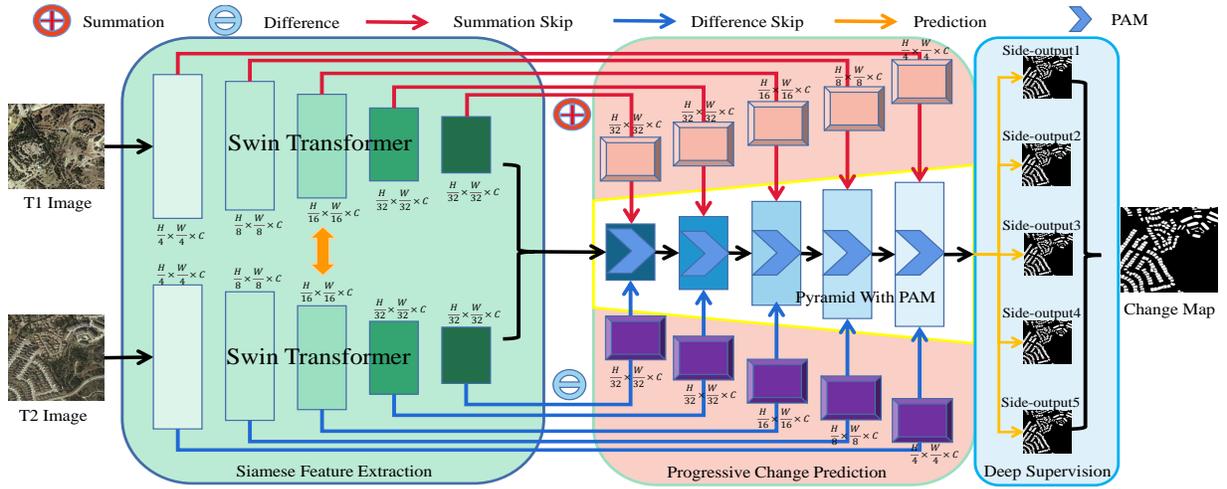
Fig. 1. The overall structure of our proposed framework (TransY-Net). Firstly, a typical Siamese Feature Extraction (SFE) network with shared Swin Transformers are utilized to extract multi-level feature maps from dual-phase remote sensing images. Then, the Deep Feature Enhancement (DFE) is introduced to highlight the change regions with summation features and difference features. Afterwards, the Progressive Change Prediction (PCP) is adopted to encode and integrate multi-level features progressively for the final change map prediction. To improve the representation ability, a pyramid structure with a Progressive Attention Module (PAM) is utilized with additional interdependencies through spatial and channel attentions. Finally, the Deep Supervison (DS) is utilized with multiple boundary-aware loss functions to train the whole framework.

However, these CD methods do not take the full abilities of Transformers in multi-level feature learning. Different from existing Transformer-based CD methods, our proposed approach handles incomplete CD regions and irregular CD boundaries. Besides, we utilize a Siamese structure to process dual-phase remote sensing images, and introduce a pyramid structure to aggregate multi-level features from Transformers for feature enhancement.

C. Feature Pyramid Methods in Remote Sensing

Multi-scale features play an important role in remote sensing image processing, including change detection. As a typical multi-scale feature fusion method, Feature Pyramid Network (FPN) [36] is first proposed for object detection from natural images, and it can aggregate multi-scale features in a coarse-to-fine manner. Recently, it also shows great successes in remote sensing tasks. For example, Yang *et al.* [37] propose a multi-scale rotation dense FPN for automatic ship detection. Li *et al.* [38] refine the FPN and multi-layer attention network for arbitrary-oriented object detection of remote sensing images. Shamsolmoali *et al.* [39] utilize a multi-patch FPN for weakly supervised object detection in optical remote sensing images. Wang *et al.* [40] enhance the FPN with deep semantic embedding for remote sensing scene classification. Gao *et al.* [41] combine multiple FPNs for end-to-end road extraction. Zhang *et al.* [42] introduce a Laplacian FPN for small object detection. Zhang *et al.* [43] combine a FPN and pixel pair matching for water-body segmentation. We note that our work is indeed inspired by the classical FPN. However, our work introduces advanced Transformers into the FPN, which can capture more long-range contextual information. Besides, our pyramid structure can improve the feature representation ability with additional inter-dependencies through spatial and channel attentions. They are different from the classical FPN.

III. PROPOSED APPROACH

As shown in Fig. 1, our proposed framework (TransY-Net) has a Y-shape structure, and includes four key components, *i.e.*, Siamese Feature Extraction (SFE), Deep Feature Enhancement (DFE), Progressive Change Prediction (PCP) and Deep Supervision (DS). By taking dual-phase images as inputs, SFE first extracts multi-level visual features through two weight-shared Swin Transformers. Then, DFE utilizes the multi-level visual features to generate summation features and difference features, which highlight the change regions with temporal information. Afterwards, by integrating all above features, PCP introduces a pyramid structure grafted with a Progressive Attention Module (PAM) for the final CD prediction. Finally, to train our proposed framework, DS is introduced to achieve the deeply-supervised learning with multiple boundary-aware loss functions for each feature level. We will elaborate on these key modules in the following subsections.

A. Siamese Feature Extraction

Following previous works, we introduce a Siamese structure to extract multi-level features from the dual-phase remote sensing images. More specifically, the Siamese structure contains two encoder branches, which share learnable weights and are used for the multi-level feature extraction of remote sensing images at temporal phase 1 (T1) and temporal phase 2 (T2), respectively. As shown in the left part of Fig. 1, we take the Swin Transformer [3] as the basic backbone of the Siamese structure, which involves five stages in total. Different from other typical Transformers [2], [23], the Swin Transformer replaces the standard Multi-Head Self-Attention (MHSA) with Window-based Multi-Head Self-Attention (W-MHSA) and Shifted Window-based Multi-Head Self-Attention (SW-MHSA), to reduce the computational complexity of the global self-attention. To improve the representation ability, the Swin Transformer also introduces the Multi-Layer Perception (MLP), LayerNorm (LN) layers and residual connections.
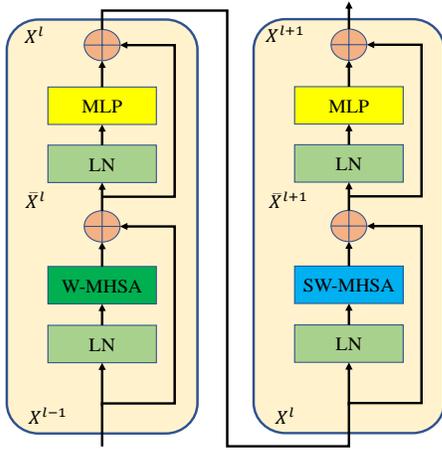
Fig. 2.   The basic structure of the used Swin Transformer block.

Fig. 2 shows the basic structure of the Swin Transformer block used in this work. Technically, the calculation formulas of all the procedures are given as follows:

$$\bar{\mathbf{X}}^l = \text{W-MHSA}(\text{LN}(\mathbf{X}^{l-1})) + \mathbf{X}^{l-1}, \qquad (1)$$

$$\mathbf{X}^l = \text{MLP}(\text{LN}(\bar{\mathbf{X}}^l)) + \bar{\mathbf{X}}^l, \qquad (2)$$

$$\bar{\mathbf{X}}^{l+1} = \text{SW-MHSA}(\text{LN}(\mathbf{X}^l)) + \mathbf{X}^l, \qquad (3)$$

$$\mathbf{X}^{l+1} = \text{MLP}(\text{LN}(\bar{\mathbf{X}}^{l+1})) + \bar{\mathbf{X}}^{l+1}, \qquad (4)$$

where $\bar{\mathbf{X}}$ is the output with the W-MHSA or SW-MHSA module and $\mathbf{X}$ is the output with the MLP module. At each stage of the original Swin Transformers, the feature resolution is halved, while the channel dimension is doubled. More specifically, the feature resolution is reduced from $(H/4) \times (W/4)$ to $(H/32) \times (W/32)$, and the channel dimension is increased from C to $8C$. In order to take advantages of global-level information, we introduce an additional Swin Transformer block to enlarge the receptive field of the feature maps. Besides, to reduce the computation, we uniformly change the channel dimension to C, and generate encoded features $[\mathbf{E}_{T1}^1, \mathbf{E}_{T1}^2, ..., \mathbf{E}_{T1}^5]$ and $[\mathbf{E}_{T2}^1, \mathbf{E}_{T2}^2, ..., \mathbf{E}_{T2}^5]$ for the T1 and T2 images, respectively. Based on the weight-shared Swin Transformers, the multi-level features can be extracted. In general, features in the high-level capture global semantic information, while features in the low-level retain local detail information. Both of them help the detection of change regions.

B. Deep Feature Enhancement

In complex scenarios, there are many visual challenges for remote sensing image CD. Thus, only depending on the above features is not enough. To highlight the change regions, we propose to enhance the multi-level visual features with feature summation and difference, as shown in the top part and bottom part of Fig. 1. Here, we note that the difference operation is a typical method for highlighting the changed regions. While the summation operation is also a useful method for feature fusion. When using the summation of two-stream features, the common information is enhanced. It is very useful for the change detection as verified in [44]. More specifically, we first

perform a point-wise feature summation and difference, then introduce a contrast feature associated to each local feature. The enhanced features can be represented as:

$$\bar{\mathbf{E}}_S^k = \text{ReLU}(\text{BN}(\text{Conv}(\mathbf{E}_{T1}^k + \mathbf{E}_{T2}^k))), \qquad (5)$$

$$\bar{\mathbf{E}}_{SC}^{k,m} = \bar{\mathbf{E}}_S^k - \text{Pool}^m(\bar{\mathbf{E}}_S^k), \qquad (6)$$

$$\mathbf{E}_S^k = [\bar{\mathbf{E}}_S^k, \bar{\mathbf{E}}_{SC}^{k,3}, ..., \bar{\mathbf{E}}_{SC}^{k,9}], \qquad (7)$$

$$\bar{\mathbf{E}}_D^k = \text{ReLU}(\text{BN}(\text{Conv}(\mathbf{E}_{T1}^k - \mathbf{E}_{T2}^k))), \qquad (8)$$

$$\bar{\mathbf{E}}_{DC}^{k,m} = \bar{\mathbf{E}}_D^k - \text{Pool}^m(\bar{\mathbf{E}}_D^k), \qquad (9)$$

$$\mathbf{E}_D^k = [\bar{\mathbf{E}}_D^k, \bar{\mathbf{E}}_{DC}^{k,3}, ..., \bar{\mathbf{E}}_{DC}^{k,9}], \qquad (10)$$

where $\mathbf{E}_S^k$ and $\mathbf{E}_D^k$ ($k = 1, 2, ..., 5$) are the enhanced features with point-wise summation and difference, respectively. ReLU is the rectified linear unit, BN is the batch normalization, Conv is a 1×1 convolution, and Pool^m is a $m \times m$ average pooling with appropriate paddings ($m \in \{3, 5, 7, 9\}$). [,] is the feature concatenation in channel. In fact, $\bar{\mathbf{E}}_{SC}^{k,m}$ and $\bar{\mathbf{E}}_{DC}^{k,m}$ capture contrast features, and can make change regions stand out from their surrounding background. Meanwhile, in most cases, keeping the original features shows better results due to the rich contextual information. Thus, we concatenate them with contrast features. Through the proposed DFE, more information of change regions and boundaries are highlighted with temporal information. Thus, the framework can make the extracted features more discriminative and obtain better CD results. We refer the readers to [45] for more insights.

C. Progressive Change Prediction

Since change regions can be any shapes and appear at any scales, we should consider the CD predictions in various cases. Inspired by the feature pyramid [36], we propose a progressive change prediction method, as shown in the middle part of Fig. 1. To improve the representation ability, a pyramid structure with a Progressive Attention Module (PAM) is utilized with additional interdependencies through spatial and channel attentions. The structure of the proposed PAM is illustrated in Fig. 3. The PAM first takes the summation features and difference features as inputs, then a Spatial-level Attention (SA) and a Channel-level Attention (CA) are jointly applied to enhance the features related to change regions. In addition, as we all know, residual connections [1] are famous, popular and efficient structures in current deep models. It alleviates the vanishing gradient problem and accelerates the training convergence. Thus, we further introduce a residual connection to improve the learning ability. The final feature map can be obtained by a 1×1 convolution. Formally, the PAM can be represented as:

$$\mathbf{F}^k = \text{ReLU}(\text{BN}(\text{Conv}([\mathbf{E}_S^k, \mathbf{E}_D^k]))), \qquad (11)$$

$$\mathbf{F}_{SA}^k = \mathbf{F}^k * \sigma(\text{Conv}(\text{SAC}(\mathbf{F}^k))), \qquad (12)$$

$$\mathbf{F}_{CA}^k = \mathbf{F}^k * \sigma(\text{Conv}(\text{GAP}(\mathbf{F}^k))), \qquad (13)$$

$$\mathbf{F}_A^k = \text{Conv}(\mathbf{F}_{SA}^k + \mathbf{F}_{CA}^k + \mathbf{F}^k), \qquad (14)$$
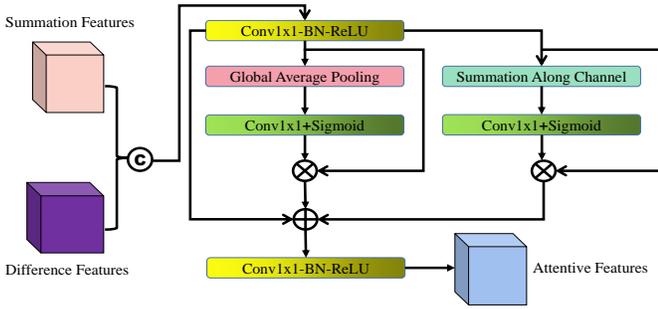
Fig. 3. The structure of our proposed Progressive Attention Module (PAM).

where σ is the Sigmoid function, SAC is the summation along the channel, and GAP is the global average pooling.

To achieve the progressive change prediction, we build the decoder pyramid grafted with a PAM as follows:

$$\mathbf{F}_P^k = \begin{cases} \mathbf{F}_A^k, & k = 5, \\ \text{UM(SwinBlock}^n(\mathbf{F}_P^{k+1}))) + \mathbf{F}_A^k, & 1 \le k < 5. \end{cases}$$ (15)

where UM is the patch unmerging block for feature upsampling [3], and SwinBlock^n is the Swin Transformer block with n layers. From the above formula, one can see that our PCP can make full use of the interdependencies within spatial and channel, and progressively aggregate multi-level features to improve the perception ability of the change regions. Here, we note that the pyramid structure is complementary to residual connections. Many previous works have already verified this fact. Thus, it is reasonable to introduce residual connections under the premise of a pyramid structure.

D. Loss Function

To optimize our framework, DS is utilized to achieve the deeply-supervised learning [46]–[48] with multiple boundary-aware loss functions for each feature level. The overall loss is defined as the summation loss over all the side-outputs and the final fusion prediction. Specifically, we first take the features of the PCP, i.e., $\mathbf{F}_P^k(k = 1, 2, ..., 5)$, and use a deconvolutional layer for the corresponding prediction $\mathbf{P}^s$ as side-outputs. Then, we concatenate them for the final fusion prediction,

$$\mathbf{P}^f = \text{Conv}[\mathbf{P}^1, ..., \mathbf{P}^S].$$ (16)

All the side-outputs and the final fusion prediction are supervised by the proposed hybrid loss:

$$\mathcal{L} = \mathcal{L}^f + \sum_{s=1}^{S} \alpha_s \mathcal{L}^s,$$ (17)

where $\mathcal{L}^f$ is the loss of the final fusion prediction and $\mathcal{L}^s$ is the loss of the s-th side-output, respectively. S denotes the total number of the side-outputs and α_s is the weight for each level loss. Our method includes five side-outputs, i.e., $S = 5$.

To obtain complete CD regions and regular CD boundaries, we define $\mathcal{L}^f$ or $\mathcal{L}^s$ as a combined loss with three terms:

$$\mathcal{L}^{f/s} = \mathcal{L}_{WBCE} + \mathcal{L}_{SSIM} + \mathcal{L}_{SIoU},$$ (18)

where $\mathcal{L}_{WBCE}$ is the weighted binary cross-entropy loss, $\mathcal{L}_{SSIM}$ is the structural similarity loss and $\mathcal{L}_{SIoU}$ is the soft intersection over union loss. The $\mathcal{L}_{WBCE}$ provides a probabilistic measure of similarity between the prediction and ground truth from a pixel-level view. The $\mathcal{L}_{SSIM}$ captures the structural information of change regions in patch-level. The $\mathcal{L}_{SIoU}$ is inspired by measuring the similarity of two sets, and yields a global similarity in CD map-level. More specifically, given the ground truth probability $g_l(\mathbf{x})$ and the estimated probability $p_l(\mathbf{x})$ at pixel $\mathbf{x}$ belong to the class l, the $\mathcal{L}_{WBCE}$ loss function is defined as:

$$\mathcal{L}_{WBCE} = -\sum_{\mathbf{x}} w(\mathbf{x}) g_l(\mathbf{x}) \log(p_l(\mathbf{x})).$$ (19)

Here, we utilize weights $w(\mathbf{x})$ to adapt the loss function to the challenges that we have encountered in CD: the class imbalance and the errors along CD boundaries. Given the frequency f_l of class l in the training data, the indicator function I, the training prediction T, and the gradient operator ∇, then the weights are defined as:

$$w(\mathbf{x}) = \sum_{l} I(T(\mathbf{x} == l)) \frac{median(\mathbf{f})}{f_l} + w_0 I(|\nabla T(\mathbf{x})| > 0),$$ (20)

where $\mathbf{f} = [f_1, ..., f_L]$ is the vector of all class frequencies. The first term models the median frequency balancing [49] and compensates for the class imbalance problem by highlighting classes with a low probability. The second term puts higher weights on the CD boundaries to emphasize on the correct detection of contours.

The $\mathcal{L}_{SSIM}$ loss considers a local neighborhood of each pixel [50]. Let $\hat{\mathbf{x}} = \{\mathbf{x}_j : j = 1, ..., N^2\}$ and $\hat{\mathbf{y}} = \{\mathbf{y}_j : j = 1, ..., N^2\}$ be the pixel values of two corresponding patches (size: $N \times N$) cropped from the prediction P and the ground truth G respectively, the $\mathcal{L}_{SSIM}$ loss is defined as:

$$\mathcal{L}_{SSIM} = 1 - \frac{(2\mu_{\mathbf{x}}\mu_{\mathbf{x}} + \epsilon)(2\sigma_{\mathbf{xy}} + \epsilon)}{(\mu_{\mathbf{x}}^2 + \mu_{\mathbf{y}}^2 + \epsilon)(\sigma_{\mathbf{x}}^2 + \sigma_{\mathbf{y}}^2 + \epsilon)},$$ (21)

where $\mu_{\mathbf{x}}$, $\mu_{\mathbf{y}}$ and $\sigma_{\mathbf{x}}$, $\sigma_{\mathbf{y}}$ are the mean and standard deviations of $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ respectively. $\sigma_{\mathbf{xy}}$ is their covariance. Here, $\epsilon = 10^{-4}$ is used to avoid dividing by zero.

In this work, one metric of interest at test time is the Intersection over Union (IoU). Thus, we also introduce the soft IoU loss [51], which is differentiable for model learning. The $\mathcal{L}_{SIoU}$ is defined as:

$$\mathcal{L}_{SIoU} = 1 - \frac{\sum_{\mathbf{x}} p_l(\mathbf{x}) g_l(\mathbf{x})}{\sum_{\mathbf{x}}[p_l(\mathbf{x}) + g_l(\mathbf{x}) - p_l(\mathbf{x}) g_l(\mathbf{x})]}.$$ (22)

When utilizing all above losses, the $\mathcal{L}_{WBCE}$ loss can relieve the class imbalance problem for change pixels, the $\mathcal{L}_{SSIM}$ loss highlights the local structure of change boundaries, and the $\mathcal{L}_{SIoU}$ loss gives more focus on the overall change regions. Thus, we can obtain better CD results and make the framework easier to optimize.

IV. EXPERIMENTS

In this section, we perform extensive experiments to verify the effectiveness of the proposed framework. We first introduce the used datasets, evaluation metrics and implementation

details. Then, we compare the proposed method with other outstanding CD methods. Finally, we perform ablation studies to verify the effectiveness of key modules with quantitative and qualitative comparisons.

A. Datasets

LEVIR-CD [52] is a public large-scale remote sensing CD dataset. It contains 637 image pairs with a 1024×1024 resolution (0.5m). We follow its default dataset split, and crop original images into small patches of size 256×256 with no overlapping. Therefore, we obtain 7120/1024/2048 pairs of image patches for training/validation/test, respectively.

WHU-CD [53] is a public building CD dataset. It contains one pair of high-resolution (0.075m) aerial images of size 32507×15354. As no pre-definite data split is widely-used, we crop the original image into small patches of size 256×256 with no overlap and randomly split them into three parts: 6096/762/762 for training/validation/test, respectively.

SYSU-CD [54] is also a public building CD dataset. It contains 20000 pairs of high-resolution (0.5m) images of size 256×256. We follow its default dataset split for experiments. There are 12000/4000/4000 pairs of image patches for training/validation/test, respectively.

Google-CD [55] is a very recent and public CD dataset. It contains 19 image pairs, originating from Google Earth Map. The image resolutions range from 1006×1168 pixels to 4936×5224 pixels. As WHU-CD, we also crop the original images into small patches of size 256×256 with no overlap and randomly split them into three parts: 2504/313/313 for training/validation/test, respectively.

In addition, we adopt two SAR image CD datasets [56] to verify the generalization of our method. These two datasets were gathered from disaster-stricken environments, which are related to flood and ice breakup situations, respectively. The Ottawa dataset is captured by the RADARSAT SAR sensor in May and August 1997 and changes are caused by floods. The size of each image is 290×350 pixels. The Sulzberger dataset is a part of Sulzberger Ice Shelf, which is provided by the European Space Agency's Envisat satellite. The size of the image is 256×256 pixels. It clearly shows the breakup of an ice shelf caused by a tsunami in March 2011.

B. Evaluation Metrics

To verify the performance of our framework and other compared methods, we follow previous works [30], [32]–[34] and utilize F1 and Intersection over Union (IoU) scores with regard to the change-class as the primary evaluation metrics. Additionally, we also report the precision and recall of the change category, Overall Accuracy (OA) and Receiver Operating Characteristic (ROC) curve. To evaluate the performance of regional boundaries, we follow previous works [57], [58] and adopt the mean Boundary Accuracy (mBA) as the metric.

C. Implementation Details

We perform experiments with the PyTorch toolbox and one NVIDIA A30 GPU. We use the mini-batch SGD algorithm to train our framework with an initial learning rate 10^{-3}, moment 0.9 and weight decay 0.0005. The batch size is set to 6. For the Siamese feature extraction backbone, we adopt the Swin Transformer pre-trained on ImageNet-22k classification task [60]. To fit the input size of the pre-trained Swin Transformer, we uniformly resize image patches to 384×384. For other layers, we randomly initialize them and set the learning rate with 10 times than the initial learning rate. We train the framework with 100 epochs. The learning rate decreases to the 1/10 of the initial learning rate at every 20 epoch. To improve the robustness, data augmentation is performed by random rotation and flipping of the input images. For the loss function in the model training, the weight parameters of each level are set equally. We release the source code at https://github.com/Drchip61/TransYNet.

D. Comparisons with State-of-the-arts

In this section, we compare the proposed method with other outstanding methods four optical and two SAR image CD datasets. These experimental results fully verify the effectiveness of our proposed framework and modules.

Quantitative Comparisons. We present the comparative results in Tab. I and Tab. II. The results clearly show that our method delivers excellent performance. More specifically, our method achieves the F1 and IoU scores of 91.90% and 83.64% on the LEVIR-CD dataset, respectively. They are much better than most of previous methods. Besides, compared with other Transformer-based methods, such as BIT [33], H-TransCD [34] and ChangeFormer [32], our method shows consistent improvements in terms of all evaluation metrics. When compared with our previous method FTN [5], the method in this paper can achieve better results in almost all metrics. On the WHU-CD dataset, our method shows significant improvements with the F1 and IoU scores of 93.38% and 87.58%, respectively. In comparison with the second-best method (FTN), our method improves the F1 and IoU scores by 1.2% and 2.1%, respectively. On the SYSU-CD dataset, our method achieves the F1 and IoU scores of 82.84% and 70.71%, respectively. The SYSU-CD dataset includes more large-scale change regions. We believe that the improvements are mainly based on the proposed DFE. On the Google-CD dataset, our method shows much better results than other compared methods. In fact, our method achieves the F1 and IoU scores of 86.04% and 75.50%, respectively. We note that the Google-CD dataset is recently proposed and it is much challenging than other three datasets. We also note that the performance of precision, recall and OA is not consistent in all methods. Our method generally achieves better recall values than most compared methods. The main reason may be that our method gives higher confidences to the change regions. To better illustrate the performance, we also present the ROC curves of some typical methods in Fig. 4. It is observed that our method achieves better results than other typical methods on the WHU-CD dataset.

Qualitative Comparisons. To illustrate the visual effect, we first display some typical CD results on the four optical image CD datasets, as shown in Fig. 5-8. From the results, one can

TABLE I

QUANTITATIVE COMPARISONS ON LEVIR-CD AND WHU-CD DATASETS. THE BEST AND THE SECOND BEST ARE IN BOLD AND UNDERLINE, RESPECTIVELY. — MEANS THE RESULTS OF CORRESPONDING METHODS ARE MISSING.

Methods	LEVIR-CD					WHU-CD				
	Pre.	Rec.	F1	IoU	OA	Pre.	Rec.	F1	IoU	OA
FC-EF [18]	86.91	80.17	83.40	71.53	98.39	71.63	67.25	69.37	53.11	97.61
FC-Siam-Diff [18]	89.53	83.31	86.31	75.92	98.67	47.33	77.66	58.81	41.66	95.63
FC-Siam-Conc [18]	91.99	76.77	83.69	71.96	98.49	60.88	73.58	66.63	49.95	97.04
BiDateNet [4]	85.65	89.98	87.76	78.19	98.52	78.28	71.59	74.79	59.73	81.92
U-Net++MSOF [17]	90.33	81.82	85.86	75.24	98.41	91.96	89.40	90.66	82.92	96.98
DTCDSCN [4]	88.53	86.83	87.67	78.05	98.77	63.92	82.30	71.95	56.19	97.42
DASNet [4]	80.76	79.53	79.91	74.65	94.32	68.14	73.03	70.50	54.41	97.29
STANet [52]	83.81	**91.00**	87.26	77.40	98.66	79.37	85.50	82.32	69.95	98.52
MSTDSNet [35]	85.52	90.84	88.10	78.73	98.56	——	——	——	——	——
IFNet [20]	**94.02**	82.93	88.13	78.77	98.87	**96.91**	73.19	83.40	71.52	98.83
SNUNet [59]	89.18	87.17	88.16	78.83	98.82	85.60	81.49	83.50	71.67	98.71
BIT [33]	89.24	89.37	89.31	80.68	98.92	86.64	81.48	83.98	72.39	98.75
H-TransCD [34]	91.45	88.72	90.06	81.92	99.00	93.85	88.73	91.22	83.85	99.24
UVACD [30]	91.90	90.70	91.30	**83.98**	**99.12**	91.45	88.72	90.06	81.92	99.00
ChangeFormer [32]	92.05	88.80	90.40	82.48	99.04	91.83	88.02	89.88	81.63	99.12
FTN [5]	92.71	89.37	91.01	83.51	99.06	93.09	91.24	92.16	85.45	99.37
Ours	92.90	89.35	**91.90**	83.64	99.07	94.68	**92.12**	**93.38**	**87.58**	**99.47**

TABLE II

QUANTITATIVE COMPARISONS ON SYSU-CD AND GOOGLE-CD DATASETS. THE BEST AND THE SECOND BEST ARE IN BOLD AND UNDERLINE, RESPECTIVELY. — MEANS THE RESULTS OF CORRESPONDING METHODS ARE MISSING.

Methods	SYSU-CD					Google-CD				
	Pre.	Rec.	F1	IoU	OA	Pre.	Rec.	F1	IoU	OA
FC-EF [18]	74.32	75.84	75.07	60.09	86.02	80.81	64.39	71.67	55.85	85.85
FC-Siam-Diff [18]	**89.13**	61.21	72.57	56.96	82.11	85.44	63.28	72.71	57.12	87.27
FC-Siam-Conc [18]	82.54	71.03	76.35	61.75	86.17	82.07	64.73	72.38	56.71	84.56
BiDateNet [4]	81.84	72.60	76.94	62.52	89.74	78.28	71.59	74.79	59.73	81.92
U-Net++MSOF [17]	81.36	75.39	78.26	62.14	86.39	91.21	57.60	70.61	54.57	95.21
DASNet [4]	68.14	70.01	69.14	60.65	80.14	71.01	44.85	54.98	37.91	90.87
STANet [52]	70.76	**85.33**	77.37	63.09	87.96	89.37	65.02	75.27	60.35	82.58
DSAMNet [20]	74.81	81.86	78.18	64.18	89.22	72.12	80.37	76.02	61.32	94.93
MSTDSNet [35]	79.91	80.76	80.33	67.13	90.67	——	——	——	——	——
SRCDNet [55]	75.54	81.06	78.20	64.21	89.34	83.74	71.49	77.13	62.77	83.18
BIT [33]	82.18	74.49	78.15	64.13	90.18	**92.04**	72.03	80.82	67.81	96.59
H-TransCD [34]	83.05	77.40	80.13	66.84	90.95	85.93	81.73	83.78	72.08	97.64
FTN [5]	86.86	76.82	81.53	68.82	91.79	86.99	**84.21**	85.58	74.79	97.92
Ours	89.09	77.42	**82.84**	**70.71**	**92.44**	87.98	84.18	**86.04**	**75.50**	**97.97**

Fig. 4. The ROC curves of some typical methods on the WHU-CD dataset.

BiT ROC curve area = 0.96
FC-EF ROC curve area = 0.84
FC-Siam-Conc ROC curve area = 0.83
FC-Siam-Diff ROC curve area = 0.79
Ours ROC curve area = 0.97
SNUNet ROC curve area = 0.95
STANet ROC curve area = 0.92

most of current methods can not detect them. However, our method can still detect them with clear boundaries, as shown in Fig. 6. In addition, when the change regions appear in complex scenes, our method can maintain the contour shape. While most of compared methods fail, as shown in Fig. 7. When distractors appear in the scene, our method can reduce the effect and correctly detect the real change regions, as shown in Fig. 8. From the above visual results, we can see that our method shows superior performance than most methods.

To further verify the visual effect, we provide more hard samples and failed results in Fig. 9. As can be seen, our method performs better than most methods (1st row). Most of current methods can not detect the two small change regions in the center, while our method can accurately localize them. Besides, we also show failed examples in the second row of Fig. 9. As can be seen, all compared methods can not detect all the change regions. However, our method shows a much more reasonable result than other methods.

To verify the generalization of our method, Fig. 10 shows

see that our method generally shows best results. For example, when change regions have multiple scales, our method can correctly identify most of the change regions, as shown in Fig. 5. When change objects cover most of the image regions,
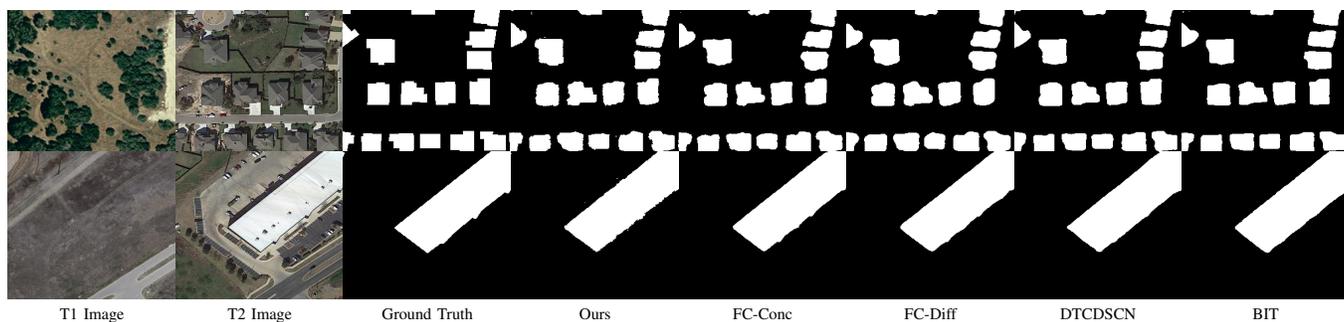
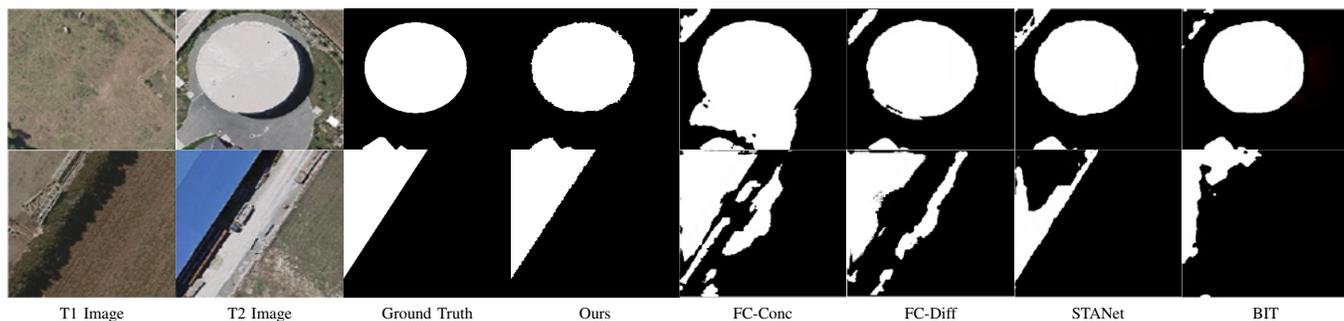Fig. 5. Comparison of typical change detection results on the LEVIR-CD dataset. Best view by zooming in.

Fig. 6. Comparison of typical change detection results on the WHU-CD dataset. Best view by zooming in.
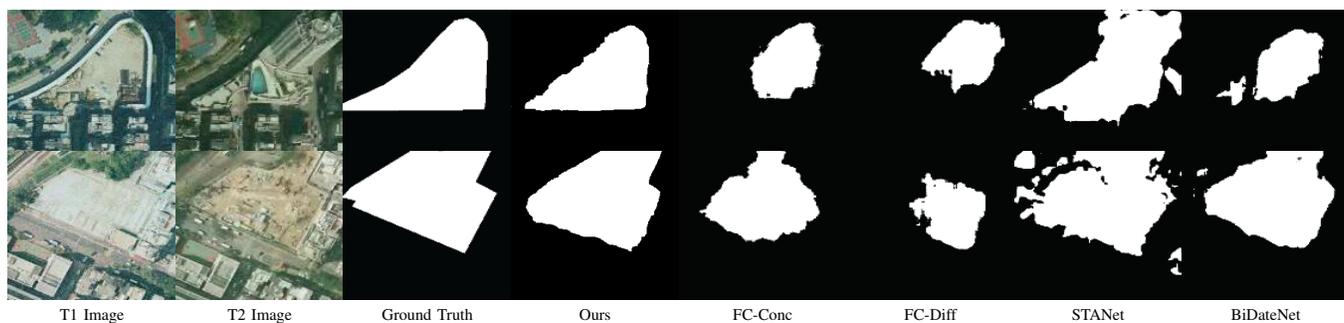

Fig. 7. Comparison of typical change detection results on the SYSU-CD dataset. Best view by zooming in.
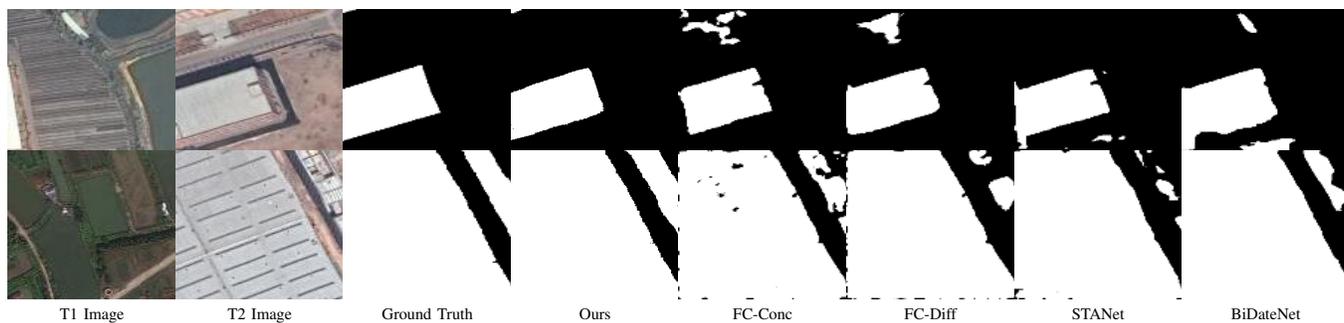

Fig. 8. Comparison of typical change detection results on the Google-CD dataset. Best view by zooming in.
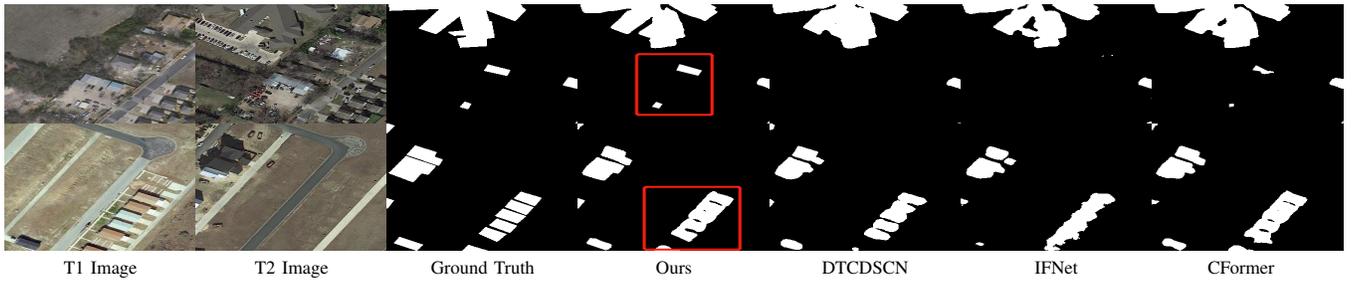
Fig. 9. Comparison of typical change detection results on more hard and failed samples. Best view by zooming in.
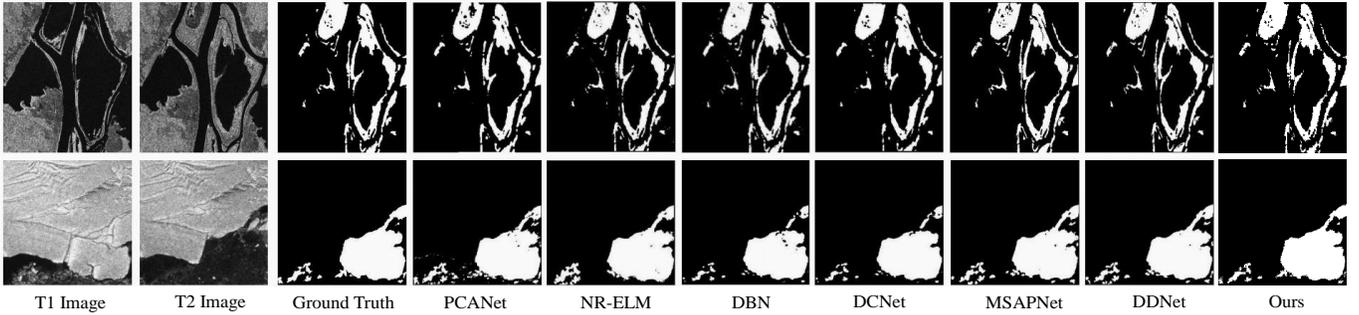

Fig. 10. Comparison of typical change detection results on the Ottawa dataset (top) and Sulzberger dataset (bottom). Images are in disaster environments. Best view by zooming in.

the change detection results on two SAR image CD datasets. The other compared methods include the PCANet [61], NR-ELM [62], DBN [63], DCNet [64], MSAPNet [65] and DDNet [66]. It can be observed that our proposed method shows better results of CD regions and boundaries. These results also demonstrate the effectiveness of our method on different types of remote sensing images and in disaster environments, including flood and tsunami.

E. More Discussions

In order to better clarify the model performance, we plot the dynamic results of our model during the training, validation and testing phases (see Fig. 11). It can be observed that our model performs better on the training data than the test data. This is reasonable and practical since most of current deep learning methods have similar trends.

Previous sections display the performance of changed regions. To evaluate the performance of regional boundaries, we list the boundary accuracy in Tab. III. It can be observed that our method shows much better results than other outstanding methods. It clearly demonstrates the effectiveness of our proposed method in improving the boundary accuracy.

We note that Transformers (including Fully Transformer Networks) are not new in current remote sensing and computer vision fields. However, there are some key differences between our work and previous methods: 1) As far as we know, our work is the earliest Transformer-based one, which explicitly handles incomplete regions and irregular boundaries for remote sensing image CD. 2) In our framework, we utilize a Siamese structure to process dual-phase remote sensing images. Besides, we introduce a pyramid structure to aggregate multi-level visual features from Transformers for
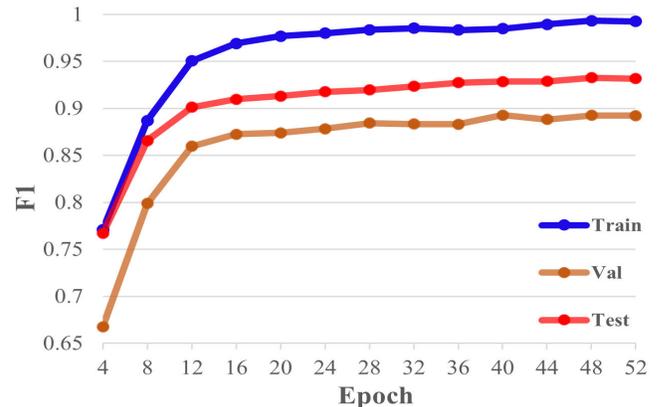

Fig. 11. Dynamic results of our model on the WHU-CD dataset.

feature enhancement. These designs are totally different from existing works, especially in [67] and [68], which mainly use a simple encoder-decoder+U-Net structure for single image feature extraction. In [67], the authors utilize a Pyramid Group Transformer (PGT) as the encoder and propose a Feature Pyramid Transformer as the decoder, which is largely based on the typical FPN structure. Meanwhile, the work in [68] simply stacks Spatial Pyramid Transformers (SPT) imitating the U-Net structure. Both of them are taking single images as inputs and using an encoder-decoder structure. While our framework utilizes a Siamese structure to process dual-phase images. 3) We utilize the deeply-supervised learning with multiple boundary-aware loss functions. These losses are very helpful for more accurate CD. These facts make our framework more convincing in the CD techniques.

TABLE III
BOUNDARY ACCURACIES WITH DIFFERENT METHODS ON LEVIR-CD.

Methods	Ours	ChangeFormer	BIT	SNUNet	STANet	FC-Diff
mBA	71.6	68.7	65.8	65.2	63.3	60.4

F. Ablation Studies

In this subsection, we perform extensive ablation studies to verify the effect of key components in our framework. The experiments are conducted on LEVIR-CD dataset. However, other datasets have similar performance trends.

Effects of different Siamese backbones. As shown in the 2-3 rows of Tab. IV, we introduce the VGGNet-16 [69] and Swin Transformer as the Siamese backbones. To ensure a fair comparison, we utilize the basic Feature Pyramid (FP) structure [36]. From the results, one can see that the performance with the Swin Transformer can be consistently improved in terms of Recall, F1, IoU and OA. The main reason is that the Swin Transformer has a better ability of modeling long-range dependency than VGGNet-16.

TABLE IV
PERFORMANCE COMPARISONS WITH DIFFERENT MODEL VARIANTS.

Models	Pre.	Rec.	F1	IoU	OA
(a) VGGNet-16+FP	91.98	82.65	87.06	77.09	98.75
(b) SwinT+FP	91.12	87.42	89.23	80.56	98.91
(c) SwinT+DFE+FP	91.73	88.43	90.05	81.89	99.00
(d) SwinT+DFE+PCP [5]	92.71	89.37	91.01	83.51	99.06
(e) SwinT+DFE+PCP	92.90	89.35	91.90	83.64	99.07

Effects of DFE. The fourth row of Tab. IV shows the effect of our proposed DFE. When compared with the $Model(b)$ $SwinT+FP$, the DFE improves the F1 value from 89.23% to 90.05%, and the IoU value from 80.56% to 81.89%, respectively. The main reason is that our DFE considers the temporal information with feature summation and difference, which highlight the change regions.

Effects of PCP. In order to better detect multi-scale change regions, we introduce the PCP, which is a pyramid structure grafted with a PAM. If we remove the PAM, the PCP will reduce a basic FP. We compare it with FP. From the results in the last two rows of Tab. IV, one can see that our PCP achieves a significant improvement in all metrics. Besides, the improved PCP in this work shows better performances. Furthermore, adding the PCP also achieves a better visual effect, in which the extracted change regions are complete and the boundaries are regular, as shown in Fig. 12.

In addition, we also introduce the Swin Transformer blocks into the PCP as shown in Eq. 11. To verify the effect of different layers, we report the results in Tab. V. From the results, one can see that the models show better results with equal layers. The best results can be achieved with $n = 4$. With more layers, the computation is larger and the performance decreases in our framework.

Effects of different losses. In this work, we introduce multiple losses to improve the CD results. To show the effects of these losses, we adopt the network structure in [5]. Tab. VI shows the results. It can be seen that using the WBCE loss can improve the F1 score from 88.75% to 90.01% and the IoU

TABLE V
PERFORMANCE COMPARISONS WITH DIFFERENT DECODER LAYERS.

Layers	Pre.	Rec.	F1	IoU	OA
(2,2,2,2)	91.18	87.00	89.04	80.24	98.90
(4,4,4,4)	91.65	88.42	90.01	81.83	99.00
(6,6,6,6)	91.70	88.30	89.96	81.76	98.99
(8,8,8,8)	91.55	88.47	89.98	81.79	98.99
(2,4,6,8)	92.13	85.71	88.80	79.86	98.89

from 79.78% to 81.83%. Using the SSIM loss achieves the F1 score of 90.11% and the IoU of 82.27%. Using the SIoU loss achieves the F1 score of 91.01% and the IoU of 83.51%.

TABLE VI
PERFORMANCE COMPARISONS WITH DIFFERENT LOSSES.

Losses	Pre.	Rec.	F1	IoU	OA
BCE	90.68	86.91	88.75	79.78	98.88
WBCE	91.65	88.42	90.01	81.83	99.00
WBCE+SSIM	91.71	88.57	90.11	82.27	99.01
WBCE+SSIM+SIoU	92.71	89.37	91.01	83.51	99.06

We also display some typical examples for the visual effects, as shown in Fig. 13. From the results, one can see that using the WBCE loss can help the model focus on the most change regions. With the SSIM loss, the framework can improve the structural information of the change regions. Using the SIoU loss can ensure the global completeness. As a result, combining all of them can achieve the best results, which proves the effectiveness of all loss terms. This fact is consistent with the quantitative results in Tab. VI.

Scaling to higher resolutions. Remote sensing images always hold large resolutions. The resolution concern is very valuable. In fact, our work can process a higher resolution with SwinT-Base/Small/Tiny. However, we adopt a low resolution (256×256), mainly considering the fairness. Most of compared methods utilize cropping for generating input images. Thus, we follow them and realize fair comparisons. Tab. VII shows the performance analysis with different resolutions on LEVIR-CD. One can see that our method can naturally scale to higher resolutions and show slightly better results.

V. CONCLUSION AND FUTURE WORK

In this work, we propose a new learning framework named TranY-Net for change detection of dual-phase remote sensing images. It improves the feature extraction from a global view and combines multi-level visual features in a pyramid manner. Technically, we first utilizes a Siamese network with the pre-trained Swin Transformers to extract long-range dependency information. Then, we introduce a pyramid structure to aggregate multi-level visual features, improving the

TABLE VII
PERFORMANCES WITH DIFFERENT INPUT RESOLUTIONS ON LEVIR-CD.

Resolution	Pre.	Rec.	F1	IoU	OA	Flops(G)
256×256	92.90	89.35	91.90	83.64	99.07	48
384×384	93.01	90.11	92.12	84.20	99.10	151
512×512	93.23	90.35	92.25	85.10	99.32	201

| T1 Image | T2 Image | Model (b) | Model (c) | Model (d) | Model (e) | Ground Truth |

Fig. 12.   Visual comparisons of predicted change maps with different models. Best view by zooming in.

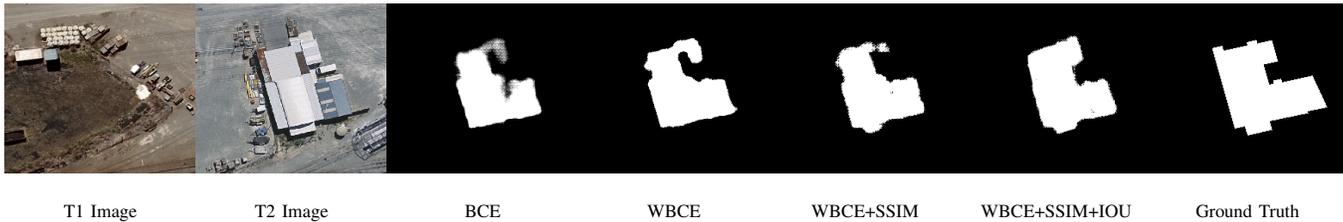| T1 Image | T2 Image | BCE | WBCE | WBCE+SSIM | WBCE+SSIM+IOU | Ground Truth |

Fig. 13.   Visual comparisons of predicted change maps with different losses. Best view by zooming in.

feature representation ability. Finally, we utilize the deeply-supervised learning with multiple loss functions for model training. Extensive experiments on four public CD benchmarks demonstrate that our proposed framework shows better performances than most state-of-the-art methods. However, our methods have some shortcomings, such as high computation, the need of image dense labeling, etc. In future works, we will explore more efficient structures of Transformers to reduce the computation. We will also develop unsupervised or weakly-supervised methods to relieve the burden of remote sensing image labeling. In addition, since our method takes dual-phase images, it can be easily used for other similar multi-modal/temporal tasks, such as RGB-D/T image segmentation, MRI-CT image fusion, video segmentation, etc. We will verify them in the computer vision field.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," pp. 770–778, 2016.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020, pp. 1–13.

[3] 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 *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[4] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 5, pp. 811–815, 2020.

[5] T. Yan, Z. Wan, and P. Zhang, "Fully transformer network for change detection of remote sensing images," in *Proceedings of Asian Conference on Computer Vision*, 2022, pp. 1–17.

[6] S. Xiaolu and C. Bo, "Change detection using change vector analysis from landsat tm images in wuhan," *Procedia Environmental Sciences*, vol. 11, pp. 238–244, 2011.

[7] Y. Tang, L. Zhang, and X. Huang, "Object-oriented change detection based on the kolmogorov–smirnov test using high-resolution multispectral imagery," *International Journal of Remote Sensing*, vol. 32, no. 20, pp. 5719–5740, 2011.

[8] L. Li, X. Li, Y. Zhang, L. Wang, and G. Ying, "Change detection for high-resolution remote sensing imagery using object-oriented change vector analysis method," in *IEEE International Geoscience and Remote Sensing Symposium*.   IEEE, 2016, pp. 2873–2876.

[9] K. Tan, Y. Zhang, X. Wang, and Y. Chen, "Object-based change detection using multiple classifiers and multi-scale uncertainty analysis," *Remote Sensing*, vol. 11, no. 3, p. 359, 2019.

[10] H. Zhang, M. Gong, P. Zhang, L. Su, and J. Shi, "Feature-level change detection using deep representation and feature change analysis for multispectral imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 11, pp. 1666–1670, 2016.

[11] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in vhr images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 3677–3693, 2019.

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

[13] D. Peng and H. Guan, "Unsupervised change detection method based on saliency analysis and convolutional neural network," *Journal of Applied Remote Sensing*, vol. 13, no. 2, p. 024512, 2019.

[14] Y. Lei, X. Liu, J. Shi, C. Lei, and J. Wang, "Multiscale superpixel segmentation with deep features for change detection," *IEEE Access*, vol. 7, pp. 36 600–36 616, 2019.

[15] N. Lv, C. Chen, T. Qiu, and A. K. Sangaiah, "Deep learning and superpixel feature extraction based on contractive autoencoder for change detection in sar images," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 12, pp. 5530–5538, 2018.

[16] P. F. Alcantarilla, S. Stent, G. Ros, R. Arroyo, and R. Gherardi, "Street-view change detection with deconvolutional networks," *Autonomous Robots*, vol. 42, no. 7, pp. 1301–1322, 2018.

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

[18] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *IEEE International Conference on Image Processing*.   IEEE, 2018, pp. 4063–4067.

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

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

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

[22] Y. Lei, D. Peng, P. Zhang, Q. Ke, and H. Li, "Hierarchical paired channel

fusion network for street scene change detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 55–67, 2020.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[24] G. Zhang, P. Zhang, J. Qi, and H. Lu, "Hat: Hierarchical aggregation transformers for person re-identification," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 516–525.

[25] X. Liu, P. Zhang, C. Yu, H. Lu, X. Qian, and X. Yang, "A video is worth three views: Trigeminal transformers for video-based person re-identification," *arXiv:2104.01745*, 2021.

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

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[28] Z. Zheng, Y. Zhong, S. Tian, A. Ma, and L. Zhang, "Changemask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 183, pp. 228–239, 2022.

[29] Z. Wang, Y. Zhang, L. Luo, and N. Wang, "Transcd: scene change detection via transformer-based architecture," *Optics Express*, vol. 29, no. 25, pp. 41 409–41 427, 2021.

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

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

[32] W. G. C. Bandara and V. M. Patel, "A transformer-based siamese network for change detection," *arXiv:2201.01293*, 2022.

[33] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

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

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

[36] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[37] X. Yang, H. Sun, K. Fu, J. Yang, X. Sun, M. Yan, and Z. Guo, "Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks," *Remote sensing*, vol. 10, no. 1, p. 132, 2018.

[38] Y. Li, Q. Huang, X. Pei, L. Jiao, and R. Shang, "Radet: Refine feature pyramid network and multi-layer attention network for arbitrary-oriented object detection of remote sensing images," *Remote Sensing*, vol. 12, no. 3, p. 389, 2020.

[39] P. Shamsolmoali, J. Chanussot, M. Zareapoor, H. Zhou, and J. Yang, "Multipatch feature pyramid network for weakly supervised object detection in optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

[40] X. Wang, S. Wang, C. Ning, and H. Zhou, "Enhanced feature pyramid network with deep semantic embedding for remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 9, pp. 7918–7932, 2021.

[41] X. Gao, X. Sun, Y. Zhang, M. Yan, G. Xu, H. Sun, J. Jiao, and K. Fu, "An end-to-end neural network for road extraction from remote sensing imagery by multiple feature pyramid network," *IEEE Access*, vol. 6, pp. 39 401–39 414, 2018.

[42] W. Zhang, L. Jiao, Y. Li, Z. Huang, and H. Wang, "Laplacian feature pyramid network for object detection in vhr optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[43] S. Chen, Y. Liu, and C. Zhang, "Water-body segmentation for multi-spectral remote sensing images by feature pyramid enhancement and pixel pair matching," *International Journal of Remote Sensing*, vol. 42, no. 13, pp. 5025–5043, 2021.

[44] H. Zhang, M. Lin, G. Yang, and L. Zhang, "Escnet: An end-to-end superpixel-enhanced change detection network for very-high-resolution remote sensing images," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 1, pp. 28–42, 2023.

[45] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6609–6617.

[46] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *IEEE International Conference on Computer Vision*, 2017, pp. 202–211.

[47] P. Zhang, L. Wang, D. Wang, H. Lu, and C. Shen, "Agile amulet: Real-time salient object detection with contextual attention," *arXiv:1802.06960*, 2018.

[48] P. Zhang, W. Liu, D. Wang, Y. Lei, H. Wang, and H. Lu, "Non-rigid object tracking via deep multi-scale spatial-temporal discriminative saliency maps," *Pattern Recognition*, vol. 100, p. 107130, 2020.

[49] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[50] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, vol. 2. Ieee, 2003, pp. 1398–1402.

[51] G. Máttyus, W. Luo, and R. Urtasun, "Deeproadmapper: Extracting road topology from aerial images," in *IEEE International Conference on Computer Vision*, 2017, pp. 3438–3446.

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

[53] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 574–586, 2018.

[54] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[55] M. Liu, Q. Shi, A. Marinoni, D. He, X. Liu, and L. Zhang, "Super-resolution-based change detection network with stacked attention module for images with different resolutions," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2021.

[56] F. Gao, X. Wang, Y. Gao, J. Dong, and S. Wang, "Sea ice change detection in sar images based on convolutional-wavelet neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 8, pp. 1240–1244, 2019.

[57] H. K. Cheng, J. Chung, Y.-W. Tai, and C.-K. Tang, "Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8890–8899.

[58] T. Shen, Y. Zhang, L. Qi, J. Kuen, X. Xie, J. Wu, Z. Lin, and J. Jia, "High quality segmentation for ultra high-resolution images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1310–1319.

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

[60] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[61] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and k-means clustering," *IEEE geoscience and remote sensing letters*, vol. 6, no. 4, pp. 772–776, 2009.

[62] F. Gao, J. Dong, B. Li, Q. Xu, and C. Xie, "Change detection from synthetic aperture radar images based on neighborhood-based ratio and extreme learning machine," *Journal of Applied Remote Sensing*, vol. 10, no. 4, pp. 046 019–046 019, 2016.

[63] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, "Change detection in synthetic aperture radar images based on deep neural networks," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 1, pp. 125–138, 2015.

[64] Y. Gao, F. Gao, J. Dong, and S. Wang, "Change detection from synthetic aperture radar images based on channel weighting-based deep cascade network," *IEEE journal of selected topics in applied earth observations and remote sensing*, vol. 12, no. 11, pp. 4517–4529, 2019.

[65] R. Wang, F. Ding, J.-W. Chen, B. Liu, J. Zhang, and L. Jiao, "Sar image change detection method via a pyramid pooling convolutional

neural network," in *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2020, pp. 312–315.

[66] X. Qu, F. Gao, J. Dong, Q. Du, and H.-C. Li, "Change detection in synthetic aperture radar images using a dual-domain network," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[67] X. He, E.-L. Tan, H. Bi, X. Zhang, S. Zhao, and B. Lei, "Fully transformer network for skin lesion analysis," *Medical Image Analysis*, vol. 77, p. 102357, 2022.

[68] S. Wu, T. Wu, F. Lin, S. Tian, and G. Guo, "Fully transformer networks for semantic image segmentation," *arXiv preprint arXiv:2106.04108*, 2021.

[69] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014.