Fully Transformer Network for Change Detection of Remote Sensing Images*

Tianyu Yan, Zifu Wan, and Pingping Zhang[0000−0003−1206−1444]

School of Artificial Intelligence, Dalian University of Technology, China
{tianyuyan2001,wanzifu2000}@gmail.com;zhpp@dlut.edu.cn

Abstract. Recently, change detection (CD) of remote sensing images have 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 learning framework named Fully Transformer Network (FTN) 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 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 interdependencies through channel attentions. Finally, to better train the 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 public CD benchmarks. For model reproduction, the source code is released at https://github.com/AI-Zhpp/FTN.

Keywords: Fully Transformer Network · Change Detection · Remote Sensing Image.

1 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 area. 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 change regions can be any shapes in complex scenarios, there are still many challenges for high-accuracy CD. In addition, remote sensing

* Supported by organization x.

image CD by handcrafted methods is time-consuming and labor-intensive, thus there is a great need for fully-automatic and highly-efficient CD.

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 deep Convolutional Neural Networks (CNN) [12,15, 17], many CD methods extract discriminative features and have demonstrated good CD performances. However, previous 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 [28]. 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 learning framework named Fully Transformer Network (FTN) 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 is a three-branch structure whose input is a dual-phase remote sensing image pair. We first utilize the advantages of Transformers [9,29,42] 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 interdependencies through 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 public CD benchmarks.

In summary, the main contributions of this work are as follow:

- We propose a novel learning framework (*i.e.*, FTN) 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 pyramid structure grafted with a Progressive Attention Module (PAM) to further improve the feature representation ability with additional interdependencies through 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 public CD benchmarks demonstrate that our framework attains better performances than most state-of-the-art methods.

2 Related Work

2.1 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 area. Before deep learning, direct classification based methods witness the great progress in CD. For example, Change Vector Analysis (CVA) [16,48] 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.* [41] propose an object-oriented CD method based on the Kolmogorov–Smirnov test. Li *et al.* [23] propose the object-oriented CVA to reduce the number of pseudo detection pixels. With multiple classifiers and multi-scale uncertainty analysis, Tan *et al.* [40] 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.* [52] utilize a Deep Belief Network (DBN) to extract deep features and represent the change regions by patch differences. Saha *et al.* [36] combine a pre-trained deep CNN and traditional CVA to generate certain change regions. Hou *et al.* [14] take the advantages of deep features and introduce the low rank analysis to improve the CD results. Peng *et al.* [33] utilize saliency detection analysis and pre-trained deep networks to achieve unsupervised CD. Since change regions may appear any places, Lei *et al.* [22] integrate Stacked Denoising AutoEncoders (SDAE) with the multi-scale superpixel segmentation to realize superpixel-based CD. Similarly, Lv *et al.* [31] utilize a Stacked Contractive AutoEncoder (SCAE) to extract temporal change features from superpixels, then adopt a clustering method to produce CD maps. Meanwhile, some methods formulate the CD task into a binary image segmentation task. Thus, CD can be finished in a supervised manner. For example, Alcantarilla *et al.* [1] 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.* [34] combine bi-temporal remote sensing images as one input, which is then fed into a modified U-Net++ [57] for CD. Daudt *et al.* [7] 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. Further more, Guo *et al.* [11] use a fully convolutional Siamese network with a contrastive loss to measure the change regions. Zhang *et al.* [49] propose a deeply-supervised image fusion network for CD. There are also some works focused on specific object CD. For example, Liu *et al.* [28] propose a dual-task constrained deep Siamese convolutional network for building CD. Jiang *et al.* [19] propose a pyramid feature-based attention-guided Siamese network for building CD. Lei *et*

al. [21] 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.

2.2 Vision Transformers for Change Detection

Recently, Transformers [42] have been applied to many computer vision tasks, such as image classification [9, 29], object detection [4], semantic segmentation [44], person re-identification [27, 51] and so on. Inspired by that, Zhang *et al.* [50] deploy a Swin Transformer [29] with a U-Net [35] structure for remote sensing image CD. Zheng *et al.* [56] design a deep Multi-task Encoder-Transformer-Decoder (METD) architecture for semantic CD. Wang *et al.* [45] 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.* [43] propose to combine a Transformer and a CNN for remote sensing image CD. Li *et al.* [24] propose an encoding-decoding hybrid framework for CD, which has the advantages of both Transformers and U-Net. Bandara *et al.* [3] 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.* [5] propose a Bitemporal Image Transformer (BIT) to efficiently and effectively model contexts within the spatial-temporal domain for CD. Ke *et al.* [20] propose a hybrid Transformer with token aggregation for remote sensing image CD. Song *et al.* [39] 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. However, these methods do not take the full abilities of Transformers in multi-level feature learning. Different from existing Transformer-based CD methods, our proposed approach improves the feature extraction from a global view and combines multi-level visual features in a pyramid manner.

3 Proposed Approach

As shown in Fig. 1, the proposed framework includes three key components, *i.e.*, Siamese Feature Extraction (SFE), Deep Feature Enhancement (DFE) and Progressive Change Prediction (PCP). By taking dual-phase remote sensing images as inputs, SFE first extracts multi-level visual features through two 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. Finally, by integrating all above features, PCP introduces a pyramid structure grafted with a Progressive Attention Module (PAM) for the final CD prediction. To train our framework, we introduce the deeply-supervised

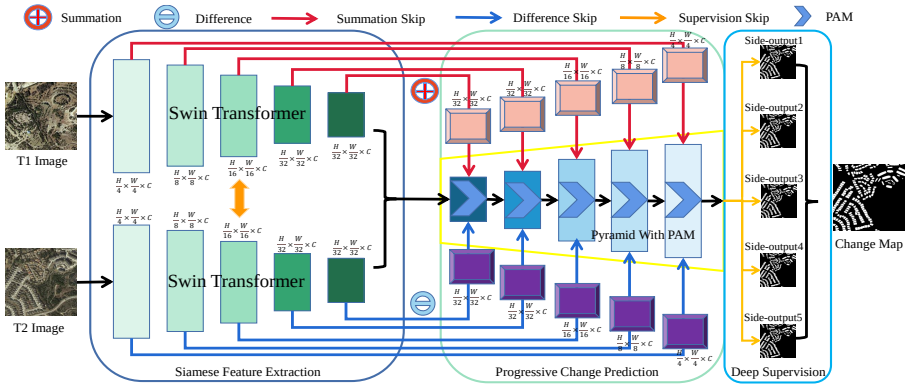

Fig. 1. The overall structure of our proposed framework.

learning with multiple boundary-aware loss functions for each feature level. We will elaborate these key modules in the following subsections.

3.1 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 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 [29] as the basic backbone of the Siamese structure, which involves five stages in total.

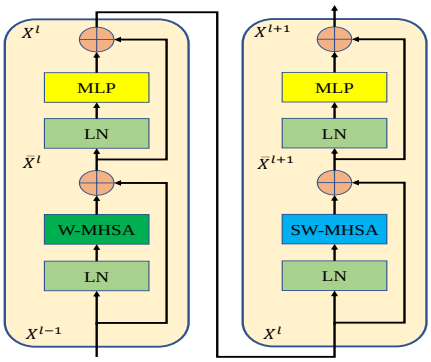

Fig. 2. The basic structure of the used Swin Transformer block.

Different from other typical Transformers [9, 42], 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 MLP, LayerNorm (LN) and residual connections. 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:

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

$$\mathbf{X}^{l} = \text{MLP}(\text{LN}(\bar{\mathbf{X}}^{l-1})) + \bar{\mathbf{X}}^{l}, \tag{2}$$

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

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

where $\bar{\mathbf{X}}$ is the output of the W-MHSA or SW-MHSA module and $\mathbf{X}$ is the output of 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 reduce 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 shared Swin Transformers, the multi-level visual features can be extracted. In general, features in the high-level capture more global semantic information, while features in the low-level retain more local detail information. Both of them help the detection of change regions.

3.2 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. More specifically, we first perform feature summation and difference, then introduce a contrast feature associated to each local feature [30]. The enhanced features can be represented as:

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

$$\mathbf{E}_{S}^{k} = [\bar{\mathbf{E}}_{S}^{k}, \bar{\mathbf{E}}_{S}^{k} - \text{Pool}(\bar{\mathbf{E}}_{S}^{k})], \tag{6}$$

$$\bar{\mathbf{E}}_{D}^{k} = \text{ReLU}(\text{BN}(\text{Conv}(\mathbf{E}_{T1}^{k} - \mathbf{E}_{T2}^{k}))), \tag{7}$$

$$\mathbf{E}_{D}^{k} = [\bar{\mathbf{E}}_{D}^{k}, \bar{\mathbf{E}}_{D}^{k} - \text{Pool}(\bar{\mathbf{E}}_{D}^{k})], \tag{8}$$

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 is a 3×3 average pooling with padding=1 and stride=1. [,] is the concatenation operation in channel. Through the proposed DFE, change regions and boundaries are highlighted with temporal information. Thus, the framework can make the extracted features more discriminative and obtain better CD results.

3.3 Progressive Change Prediction

Since change regions can be any shapes and appear in any scales, we should consider the CD predictions at various cases. Inspired by the feature pyramid [25], we propose a progressive change prediction, 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 channel attentions. The structure of the proposed PAM is illustrated in Fig. 3. It first takes the summation features and difference features as inputs, then a channel-level attention is applied to enhance the features related to change regions. Besides, we also 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:

$$\mathbf{F}^k = \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}([\mathbf{E}_S^k, \mathbf{E}_D^k]))), \tag{9}$$

$$\mathbf{F}_A^k = \mathbf{F}^k * \sigma(\mathrm{Conv}(\mathrm{GAP}(\mathbf{F}^k))) + \mathbf{F}^k, \tag{10}$$

where σ is the Sigmoid function and GAP is the global average pooling.

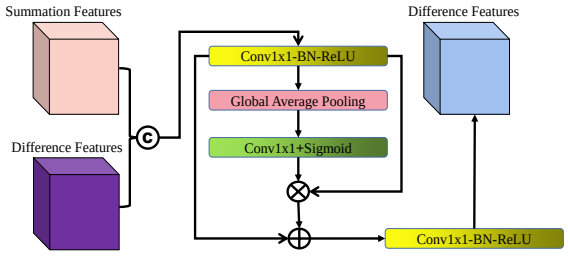

Fig. 3. The structure of our proposed Progressive Attention Module (PAM).

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

$$\mathbf{F}_P^k = \begin{cases} \mathbf{F}_A^k, & k = 5, \\ \mathrm{UM}(\mathrm{SwinBlock}^n(\mathbf{F}_P^{k+1})) + \mathbf{F}_A^k, & 1 \le k < 5. \end{cases} \tag{11}$$

where UM is the patch unmerging block used in Swin Transformers for upsampling, and $\mathrm{SwinBlock}^n$ 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 channels, and can progressively aggregate multi-level visual features to improve the perception ability of the change regions.

3.4 Loss Function

To optimize our framework, we adopt the deeply-supervised learning [53–55] with multiple boundary-aware loss functions for each feature level. The overall loss is defined as the summation over all side-outputs and the final fusion prediction:

$$\mathcal{L} = \mathcal{L}^f + \sum_{s=1}^{S} \alpha_s \mathcal{L}^s \tag{12}$$

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. In this work, 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:

$$\mathcal{L}^{f/s} = \mathcal{L}_{WBCE} + \mathcal{L}_{SSIM} + \mathcal{L}_{SIoU}, \tag{13}$$

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 map-level. More specifically, given the ground truth probability $g_l(\mathbf{x})$ and the estimated probability $p_l(\mathbf{x})$ at pixel $\mathbf{x}$ to belong to the class l, the $\mathcal{L}_{WBCE}$ loss function is

$$\mathcal{L}_{WBCE} = -\sum_{\mathbf{x}} w(\mathbf{x}) g_l(\mathbf{x}) \log(p_l(\mathbf{x})). \tag{14}$$

Here, we utilize weights $w(\mathbf{x})$ to handle challenges appeared 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 P, and the gradient operator ∇, weights are defined as:

$$w(\mathbf{x}) = \sum_{l} I(P(\mathbf{x} == l)) \frac{median(\mathbf{f})}{f_l} + w_0 I(|\nabla P(\mathbf{x})| > 0), \tag{15}$$

where $\mathbf{f} = [f_1, ..., f_L]$ is the vector of all class frequencies. The first term models median frequency balancing [2] to handle the class imbalance problem by highlighting classes with low probability. The second term assigns higher weights on the CD boundaries to emphasize on the correct prediction of boundaries.

The $\mathcal{L}_{SSIM}$ loss considers a local neighborhood of each pixel [46]. 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:

$$\mathcal{L}_{SSIM} = 1 - \frac{(2\mu_{\mathbf{x}}\mu_{\mathbf{x}} + \epsilon)(2\sigma_{\mathbf{xy}} + \epsilon)}{(\mu_{\mathbf{x}}^2 + \mu_{\mathbf{y}}^2 + \epsilon)(\sigma_{\mathbf{x}}^2 + \sigma_{\mathbf{y}}^2 + \epsilon)}, \tag{16}$$

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. $\epsilon = 10^{-4}$ 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 [32], which is differentiable for learning. The $\mathcal{L}_{SIoU}$ is defined as:

$$\mathcal{L}_{SIoU} = 1 - \frac{\sum_{\mathbf{x}} p_l(\mathbf{x}) g_l(\mathbf{x})}{\sum_{\mathbf{x}} [p_l(\mathbf{x}) + g_l(\mathbf{x}) - p_l(\mathbf{x}) g_l(\mathbf{x})]}. \tag{17}$$

When utilizing all above losses, the $\mathcal{L}_{WBCE}$ loss can relieve the 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 change regions. Thus, we can obtain better CD results and make the framework easier to optimize.

4 Experiments

4.1 Datasets

LEVIR-CD [6] is a public large-scale CD dataset. It contains 637 remote sensing 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 [18] is a public building CD dataset. It contains one pair of high-resolution (0.075m) aerial images of size 32507×15354. As no definite data split is widely-used, we crop the original image into small patches of size 256×256 with no overlap and randomly split it into three parts: 6096/762/762 for training/validation/test, respectively.

SYSU-CD [37] 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 [26] is a very recent and public CD dataset. It contains 19 image pairs, originating from Google Earth Map. The image resolutions are ranging from 1006×1168 pixels to 4936×5224 pixels. We crop the images into small patches of size 256×256 with no overlap and randomly split it into three parts: 2504/313/313 for training/validation/test, respectively.

4.2 Evaluation Metrics

To verify the performance, we follow previous works [3,49] and mainly 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 and overall accuracy (OA).

4.3 Implementation Details

We perform experiments with the public MindSpore toolbox and one NVIDIA A30 GPU. We used 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 [8]. 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. For model reproduction, the source code is released at https://github.com/AI-Zhpp/FTN.

4.4 Comparisons with Sate-of-the-arts

In this section, we compare the proposed method with other outstanding methods on four public CD datasets. The experimental results fully verify the effectiveness of our proposed method.

Table 1. Quantitative comparisons on LEVIR-CD and WHU-CD datasets.

	LEVIR-CD					WHU-CD				
Methods	Pre.	Rec.	F1	IoU	OA	Pre.	Rec.	F1	IoU	OA
FC-EF [7]	86.91	80.17	83.40	71.53	98.39	71.63	67.25	69.37	53.11	97.61
FC-Siam-Diff [7]	89.53	83.31	86.31	75.92	98.67	47.33	77.66	58.81	41.66	95.63
FC-Siam-Conc [7]	91.99	76.77	83.69	71.96	98.49	60.88	73.58	66.63	49.95	97.04
BiDateNet [28]	85.65	89.98	87.76	78.19	98.52	78.28	71.59	74.79	59.73	81.92
U-Net++MSOF [34]	90.33	81.82	85.86	75.24	98.41	91.96	89.40	90.66	82.92	96.98
DTCDSCN [28]	88.53	86.83	87.67	78.05	98.77	63.92	82.30	71.95	56.19	97.42
DASNet [28]	80.76	79.53	79.91	74.65	94.32	68.14	73.03	70.50	54.41	97.29
STANet [6]	83.81	**91.00**	87.26	77.40	98.66	79.37	85.50	82.32	69.95	98.52
MSTDSNet [39]	85.52	90.84	88.10	78.73	98.56	—	—	—	—	—
IFNet [49]	**94.02**	82.93	88.13	78.77	98.87	**96.91**	73.19	83.40	71.52	98.83
SNUNet [10]	89.18	87.17	88.16	78.83	98.82	85.60	81.49	83.50	71.67	98.71
BIT [5]	89.24	89.37	89.31	80.68	98.92	86.64	81.48	83.98	72.39	98.75
H-TransCD [20]	91.45	88.72	90.06	81.92	99.00	93.85	88.73	91.22	83.85	99.24
ChangeFormer [3]	92.05	88.80	90.40	82.48	99.04	91.83	88.02	89.88	81.63	99.12
Ours	92.71	89.37	**91.01**	**83.51**	**99.06**	93.09	**91.24**	**92.16**	**85.45**	**99.37**

Quantitative Comparisons. We present the comparative results in Tab. 1 and Tab. 2. The results show that our method delivers excellent performance. More specifically, our method achieves the best F1 and IoU values of 91.01% and 83.51% on the LEVIR-CD dataset, respectively. They are much better than previous methods. Besides, compared with other Transformer-based methods, such

as BIT [5], H-TransCD [20] and ChangeFormer [3], our method shows consistent improvements in terms of all evaluation metrics. On the WHU-CD dataset, our method shows significant improvement with the F1 and IoU values of 92.16% and 85.45%, respectively. Compared with the second-best method, our method improves the F1 and IoU values by 0.9% and 1.6%, respectively. On the SYSU-CD dataset, our method achieves the F1 and IoU values of 81.53% and 68.82%, respectively. The SYSU-CD dataset includes more large-scale change regions. We think the improvements are mainly based on the proposed DFE. On the Google-CD dataset, our method shows much better results than compared methods. In fact, our method achieves the F1 and IoU values of 85.58% and 74.79%, 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 achieve better recall values than most methods. The main reason may be that our method gives higher confidences to the change regions.

Table 2. Quantitative comparisons on SYSU-CD and Google-CD datasets.

Methods	SYSU-CD					Google-CD				
	Pre.	Rec.	F1	IoU	OA	Pre.	Rec.	F1	IoU	OA
FC-EF [7]	74.32	75.84	75.07	60.09	86.02	80.81	64.39	71.67	55.85	85.85
FC-Siam-Diff [7]	**89.13**	61.21	72.57	56.96	82.11	85.44	63.28	72.71	57.12	87.27
FC-Siam-Conc [7]	82.54	71.03	76.35	61.75	86.17	82.07	64.73	72.38	56.71	84.56
BiDateNet [28]	81.84	72.60	76.94	62.52	89.74	78.28	71.59	74.79	59.73	81.92
U-Net++MSOF [34]	81.36	75.39	78.26	62.14	86.39	91.21	57.60	70.61	54.57	95.21
DASNet [28]	68.14	70.01	69.14	60.65	80.14	71.01	44.85	54.98	37.91	90.87
STANet [6]	70.76	**85.33**	77.37	63.09	87.96	**89.37**	65.02	75.27	60.35	82.58
DSAMNet [49]	74.81	81.86	78.18	64.18	89.22	72.12	80.37	76.02	61.32	94.93
MSTDSNet [39]	79.91	80.76	80.33	67.13	90.67	—	—	—	—	—
SRCDNet [26]	75.54	81.06	78.20	64.21	89.34	83.74	71.49	77.13	62.77	83.18
BIT [5]	82.18	74.49	78.15	64.13	90.18	92.04	72.03	80.82	67.81	96.59
H-TransCD [20]	83.05	77.40	80.13	66.84	90.95	85.93	81.73	83.78	72.08	97.64
Ours	86.86	76.82	**81.53**	**68.82**	**91.79**	86.99	**84.21**	**85.58**	**74.79**	**97.92**

Qualitative Comparisons. To illustrate the visual effect, we display some typical CD results on the four datasets, as shown in Fig. 4. From the results, we can see that our method generally shows best CD results. For example, when change regions have multiple scales, our method can correctly identify most of them, as shown in the first row. When change objects cover most of the image regions, most of current methods can not detect them. However, our method can still detect them with clear boundaries, as shown in the second row. In addition, when change regions appear in complex scenes, our method can maintain the contour shape. While most of compared methods fail, as shown in the third row. When distractors appear, our method can reduce the effect and correctly detect change regions, as shown in the fourth row. From these visual results, we can see that our method shows superior performance than most methods.

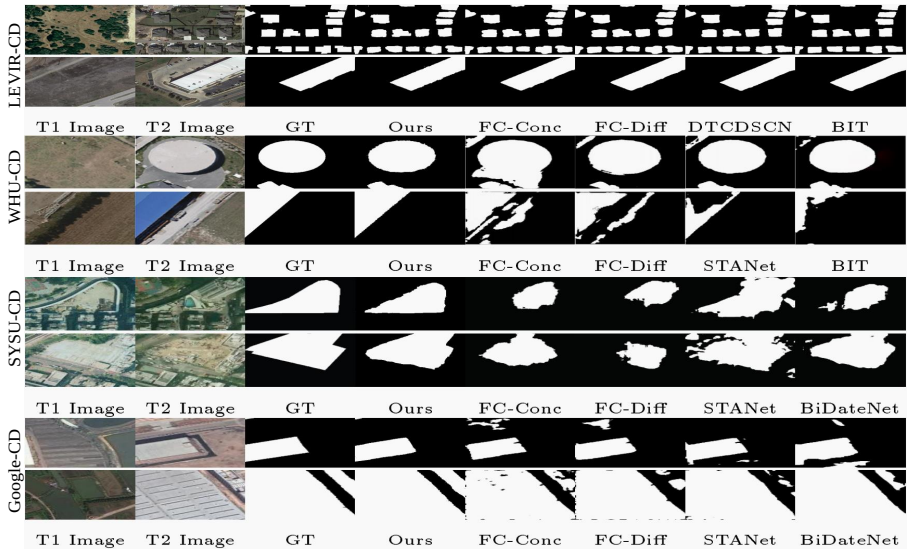

Fig. 4. Comparison of typical change detection results on four CD datasets.

To further verify the effectiveness, we provide more hard samples in Fig. 5. 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. 5. As can be seen, all compared methods can not detect all the change regions. However, our method shows more reasonable results.

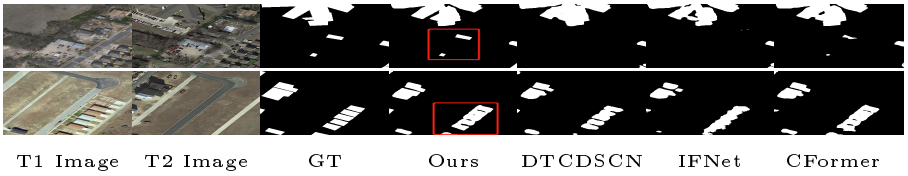

Fig. 5. Comparison of typical change detection results on hard and failed samples.

4.5 Ablation study

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. 3, we introduce the VGGNet-16 [38] and Swin Transformer as Siamese

backbones. To ensure a fair comparison, we utilize the basic Feature Pyramid (FP) structure [25]. 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 3. Performance comparisons with different model variants on LEVIR-CD.

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	92.71	89.37	91.01	83.51	99.06

Effects of DFE. The fourth row of Tab. 3 shows the effect of our proposed DFE. When compared with the $Model(b)$ $SwinT + FP$, 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 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. We compare it with FP. From the results in the last row of Tab. 3, one can see that our PCP achieves a significant improvement in all metrics. 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. 6.

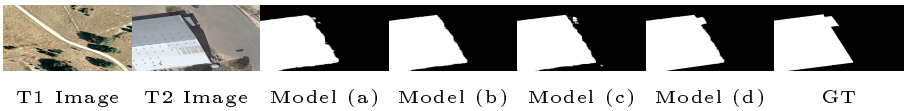

Fig. 6. Visual comparisons of predicted change maps with different models.

In addition, we also introduce the Swin Transformer blocks in the PCP as shown in Eq. 11. To verify the effect of different layers, we report the results in Tab. 4. From the results, we 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 loss functions to improve the CD results. Tab. 5 shows the effects of these losses. It can be seen that using the WBCE loss can improve the F1 value from 88.75% to 90.01% and the IoU from 79.78% to 81.83%. Using the SSIM loss achieves the F1 value of 90.11% and the IoU of 82.27%. Using the SIoU loss achieves the F1 value of 91.01% and the IoU of 83.51%. In fact, combining all of them can achieve the best results, which prove the effectiveness of all loss terms.

Table 4. Performance comparisons with different decoder layers on LEVIR-CD.

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

Table 5. Performance comparisons with different losses on LEVIR-CD.

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

More structure discussions. There are some key differences between our work and previous fully Transformer structures: The works in [13,47] are taking single images as inputs and using an encoder-decoder structure. However, our framework utilizes a Siamese structure to process dual-phase images. In order to fuse features from two encoder streams, we propose a pyramid structure grafted with a PAM for the final CD prediction. Thus, apart from the input difference, our work progressively aggregates multi-level features for feature enhancement.

5 Conclusion

In this work, we propose a new learning framework named FTN for change detection of dual-phase remote sensing images. 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 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. In future works, we will explore more efficient structures of Transformers to reduce the computation and develop unsupervised or weakly-supervised methods to relieve the burden of remote sensing image labeling.

Acknowledgements This work is partly sponsored by CAAI-Huawei MindSpore Open Fund (No.CAAIXSJLJJ-2021-067A), the College Students' Innovative Entrepreneurial Training Plan Program (No. 20221014141123) and the Fundamental Research Funds for the Central Universities (No. DUT20RC(3)083).

References

1. Alcantarilla, P.F., Stent, S., Ros, G., Arroyo, R., Gherardi, R.: Street-view change detection with deconvolutional networks. Autonomous Robots **42**(7), 1301–1322 (2018)
2. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(12), 2481–2495 (2017)
3. Bandara, W.G.C., Patel, V.M.: A transformer-based siamese network for change detection. arXiv:2201.01293 (2022)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision. pp. 213–229 (2020)
5. Chen, H., Qi, Z., Shi, Z.: Remote sensing image change detection with transformers. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–14 (2021)
6. Chen, H., Shi, Z.: A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. Remote Sensing **12**(10), 1662 (2020)
7. Daudt, R.C., Le Saux, B., Boulch, A.: Fully convolutional siamese networks for change detection. In: IEEE International Conference on Image Processing. pp. 4063–4067. IEEE (2018)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations. pp. 1–13 (2020)
10. Fang, S., Li, K., Shao, J., Li, Z.: Snunet-cd: A densely connected siamese network for change detection of vhr images. IEEE Geoscience and Remote Sensing Letters **19**, 1–5 (2021)
11. Guo, E., Fu, X., Zhu, J., Deng, M., Liu, Y., Zhu, Q., Li, H.: Learning to measure change: Fully convolutional siamese metric networks for scene change detection. arXiv:1810.09111 (2018)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
13. He, X., Tan, E.L., Bi, H., Zhang, X., Zhao, S., Lei, B.: Fully transformer network for skin lesion analysis. Medical Image Analysis **77**, 102357 (2022)
14. Hou, B., Wang, Y., Liu, Q.: Change detection based on deep features and low rank. IEEE Geoscience and Remote Sensing Letters **14**(12), 2418–2422 (2017)
15. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708 (2017)
16. Huo, C., Zhou, Z., Lu, H., Pan, C., Chen, K.: Fast object-level change detection for vhr images. IEEE Geoscience and Remote Sensing Letters **7**(1), 118–122 (2009)
17. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. pp. 448–456 (2015)
18. Ji, S., Wei, S., Lu, M.: Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Transactions on Geoscience and Remote Sensing **57**(1), 574–586 (2018)

19. Jiang, H., Hu, X., Li, K., Zhang, J., Gong, J., Zhang, M.: Pga-siamnet: Pyramid feature-based attention-guided siamese network for remote sensing orthoimagery building change detection. Remote Sensing **12**(3), 484 (2020)
20. Ke, Q., Zhang, P.: Hybrid-transcd: A hybrid transformer remote sensing image change detection network via token aggregation. ISPRS International Journal of Geo-Information **11**(4), 263 (2022)
21. Lei, Y., Peng, D., Zhang, P., Ke, Q., Li, H.: Hierarchical paired channel fusion network for street scene change detection. IEEE Transactions on Image Processing **30**, 55–67 (2020)
22. Lei, Y., Liu, X., Shi, J., Lei, C., Wang, J.: Multiscale superpixel segmentation with deep features for change detection. IEEE Access **7**, 36600–36616 (2019)
23. Li, L., Li, X., Zhang, Y., Wang, L., Ying, G.: Change detection for high-resolution remote sensing imagery using object-oriented change vector analysis method. In: IEEE International Geoscience and Remote Sensing Symposium. pp. 2873–2876. IEEE (2016)
24. Li, Q., Zhong, R., Du, X., Du, Y.: Transunetcd: A hybrid transformer network for change detection in optical remote-sensing images. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–19 (2022)
25. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2117–2125 (2017)
26. Liu, M., Shi, Q., Marinoni, A., He, D., Liu, X., Zhang, L.: Super-resolution-based change detection network with stacked attention module for images with different resolutions. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–18 (2021)
27. Liu, X., Zhang, P., Yu, C., Lu, H., Qian, X., Yang, X.: A video is worth three views: Trigeminal transformers for video-based person re-identification. arXiv:2104.01745 (2021)
28. Liu, Y., Pang, C., Zhan, Z., Zhang, X., Yang, X.: Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model. IEEE Geoscience and Remote Sensing Letters **18**(5), 811–815 (2020)
29. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
30. Luo, Z., Mishra, A., Achkar, A., Eichel, J., Li, S., Jodoin, P.M.: Non-local deep features for salient object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 6609–6617 (2017)
31. Lv, N., Chen, C., Qiu, T., Sangaiah, A.K.: Deep learning and superpixel feature extraction based on contractive autoencoder for change detection in sar images. IEEE Transactions on Industrial Informatics **14**(12), 5530–5538 (2018)
32. Máttyus, G., Luo, W., Urtasun, R.: Deeproadmapper: Extracting road topology from aerial images. In: IEEE International Conference on Computer Vision. pp. 3438–3446 (2017)
33. Peng, D., Guan, H.: Unsupervised change detection method based on saliency analysis and convolutional neural network. Journal of Applied Remote Sensing **13**(2), 024512 (2019)
34. Peng, D., Zhang, Y., Guan, H.: End-to-end change detection for high resolution satellite images using improved unet++. Remote Sensing **11**(11), 1382 (2019)
35. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)

36. Saha, S., Bovolo, F., Bruzzone, L.: Unsupervised deep change vector analysis for multiple-change detection in vhr images. IEEE Transactions on Geoscience and Remote Sensing **57**(6), 3677–3693 (2019)
37. Shi, Q., Liu, M., Li, S., Liu, X., Wang, F., Zhang, L.: 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 **60**, 1–16 (2021)
38. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
39. Song, F., Zhang, S., Lei, T., Song, Y., Peng, Z.: Mstdsnet-cd: Multiscale swin transformer and deeply supervised network for change detection of the fast-growing urban regions. IEEE Geoscience and Remote Sensing Letters **19**, 1–5 (2022)
40. Tan, K., Zhang, Y., Wang, X., Chen, Y.: Object-based change detection using multiple classifiers and multi-scale uncertainty analysis. Remote Sensing **11**(3), 359 (2019)
41. Tang, Y., Zhang, L., Huang, X.: Object-oriented change detection based on the kolmogorov–smirnov test using high-resolution multispectral imagery. International Journal of Remote Sensing **32**(20), 5719–5740 (2011)
42. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
43. Wang, G., Li, B., Zhang, T., Zhang, S.: A network combining a transformer and a convolutional neural network for remote sensing image change detection. Remote Sensing **14**(9), 2228 (2022)
44. Wang, Y., Xu, Z., Wang, X., Shen, C., Cheng, B., Shen, H., Xia, H.: End-to-end video instance segmentation with transformers. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 8741–8750 (2021)
45. Wang, Z., Zhang, Y., Luo, L., Wang, N.: Transcd: scene change detection via transformer-based architecture. Optics Express **29**(25), 41409–41427 (2021)
46. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers. vol. 2, pp. 1398–1402. Ieee (2003)
47. Wu, S., Wu, T., Lin, F., Tian, S., Guo, G.: Fully transformer networks for semantic image segmentation. arXiv:2106.04108 (2021)
48. Xiaolu, S., Bo, C.: Change detection using change vector analysis from landsat tm images in wuhan. Procedia Environmental Sciences **11**, 238–244 (2011)
49. Zhang, C., Yue, P., Tapete, D., Jiang, L., Shangguan, B., Huang, L., Liu, G.: A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing **166**, 183–200 (2020)
50. Zhang, C., Wang, L., Cheng, S., Li, Y.: Swinsunet: Pure transformer network for remote sensing image change detection. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–13 (2022)
51. Zhang, G., Zhang, P., Qi, J., Lu, H.: Hat: Hierarchical aggregation transformers for person re-identification. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 516–525 (2021)
52. Zhang, H., Gong, M., Zhang, P., Su, L., Shi, J.: Feature-level change detection using deep representation and feature change analysis for multispectral imagery. IEEE Geoscience and Remote Sensing Letters **13**(11), 1666–1670 (2016)
53. Zhang, P., Liu, W., Wang, D., Lei, Y., Wang, H., Lu, H.: Non-rigid object tracking via deep multi-scale spatial-temporal discriminative saliency maps. Pattern Recognition **100**, 107130 (2020)

54. Zhang, P., Wang, D., Lu, H., Wang, H., Ruan, X.: Amulet: Aggregating multi-level convolutional features for salient object detection. In: IEEE International Conference on Computer Vision. pp. 202–211 (2017)
55. Zhang, P., Wang, L., Wang, D., Lu, H., Shen, C.: Agile amulet: Real-time salient object detection with contextual attention. arXiv:1802.06960 (2018)
56. Zheng, Z., Zhong, Y., Tian, S., Ma, A., Zhang, L.: Changemask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection. ISPRS Journal of Photogrammetry and Remote Sensing **183**, 228–239 (2022)
57. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 3–11. Springer (2018)