REMOTE SENSING IMAGES CHANGE DETECTION USING THE SIAMESE NETWORK COMBINED WITH PURE SWIN TRANSFORMER

Xu SONG[1,2], Xinyu TONG*[1], Asif Iqbal HAJAMYDEEN*[2]

While Transformers have become prevalent in detecting changes in remote sensing imagery, several challenges hinder their broader adoption in the field, such as missing detection, low precision of contour boundary detection and the complexity of calculation in the data processing. Addressing these challenges, our research introduces the SiamFormer architecture, which employs a layered Swin Transformer encoder coupled with a cascaded decoder for feature fusion, specifically designed for change detection within remote sensing imagery. First, our approach integrates the Siamese network framework with a pure Swin Transformer, crafting a decoder with a hierarchical layout to bolster its capabilities for pixel-level change detection in remote sensing images. Second, we improve the interconnection status between decoder layers by top-down cascade paths and dense cascades to produce high-quality high-resolution image semantic feature change outputs. Furthermore, to mitigate the loss of edge details in change objects caused by high-dimensional downsampling, we implement a convolutional decoding classifier. This classifier efficiently reduces the channel dimensions of the merged change feature map to the bare minimum. Our experimental analysis, conducted on the CDD and LEVIR-CD datasets, demonstrates that our proposed methodology outperforms existing change detection techniques for remote sensing imagery in terms of effectiveness.

Key words: change detection; high-resolution remote sensing; Swin Transformer; Self-Attention mechanism; Siamese network

1. Introduction

Change detection (CD) is a distinct technological approach used to identify alterations in terrestrial objects by analyzing and juxtaposing multiple images of the same geographic area taken at various times. Currently, this technology has been extensively used in forest land cover, urban planning, and environmental monitoring.

Recently, deep learning (DL) has exhibited strong feature extraction ability in various computer vision (CV) tasks. The DL-based CD can automatically extract the deep change features of remote sensing images for image segmentation, which

[1] School of Computer and Information Engineering, Anyang Normal University, Anyang 455000, China
[2] School of Graduate Studies, Management and Science University, Shah Alam 40100, Malaysia,
* Corresponding authors' e-mail: tonxycs@yeah.net, asif@msu.edu.my

significantly reduces artificial feature engineering and help complete large-scale CD tasks more robustly. Daudt et al. [1] introduced the concepts of fully convolutional early fusion (FC-EF) and FC-Siam-Conc, incorporating jump connections akin to those found in UNet [2], to precisely identify regions of change. In a related development, Peng et al. [3] formulated the U-Net++MSOF network model for CD, which integrates deep supervision with a dense connection strategy to refine the edge delineation of areas undergoing change. Chen et al. [4] studied the same field and captured the spatiotemporal dependence of images at different scales through spatiotemporal attention mechanisms based on a multi-scale spatiotemporal attention CD model of Siamese structures (STA-Net). Inspired by the Transformer's design in NLP, researchers have developed a variety of deep neural network models for image-related tasks, such as object detection and segmentation. Notable examples include the Visual Transformer (ViT) [5], the shift window-based Swin Transformer [6], and SegFormer [7]. These Transformer-based models offer an advantage over traditional deep convolutional methods by providing a broader effective receptive field (ERF) and a superior capability to understand the contextual relationships between any two pixels in an image. However, they are not immune to the challenges posed by variations in ground object shapes, scene complexity, imaging conditions, and alignment inaccuracies, which can lead to omissions and false positives. While, although the Transformer architecture excels in capturing global context, the multi-head self-attention (MSA) mechanism within it involves a vast number of tokens, resulting in extensive calculations for global self-attention.

In response, we developed the SiamFormer, a network that combines Siamese configurations with hierarchical Swin Transformer encoding frameworks, aiming to streamline data processing and boost the capacity for modeling global receptive fields and distant context relationships. Furthermore, we devised an encoder for hierarchical feature extraction and a decoder for cascading feature fusion, both rooted in the Swin Transformer architecture. This design is intended to advance the extraction of global context and the modeling of extensive contextual dependencies for pixel-level CD in remote sensing imagery. Meanwhile we employed cascading fusion techniques to integrate features across scales and generate detailed maps predicting changes.

2 Related Work

2.1 Siamese Network. Siamese networks are also called twin networks as shown in Fig. 1, and their major characteristic is the ability to share weights. Using the process of CD as an illustration, when two akin images (Input1 and Input2) are fed into the identical network for training purposes, the calculation of loss hinges on their discrepancy. After backpropagation training, the output feature map of both

images can be predicted by the Siamese network. In this case, when feature maps are very similar, they would subsequently represent a smaller eigenvalue difference of the corresponding vector matrix. Building on this premise, the feature map produced by the Siamese network is applicable for tasks such as target detection and identification.

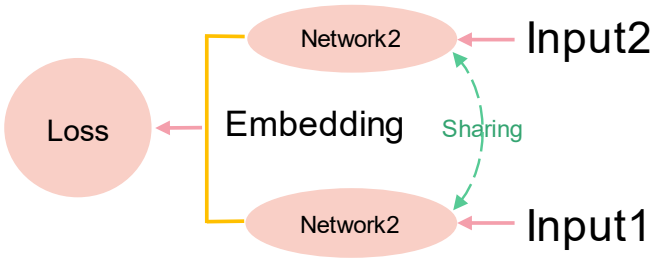

Fig. 1. Siamese network

2.2 Transformers in CD. Initially introduced for sequence processing tasks in machine translation within NLP [8], the Transformer architecture has garnered attention for its success in recent years in image feature extraction methods within CV tasks. Notably, the pioneering ViT network has achieved a favorable balance between speed and precision in image recognition. Several researchers have explored the integration of CD with CV tasks and put forward a novel hierarchical visual Transformer named Swin Transformer, which has demonstrated remarkable efficacy and efficiency. Serving as a visual backbone, the Swin Transformer has showcased superior performance across various visual tasks such as image classification, target detection, and semantic segmentation, owing to its innovative Multi-Scale Attention (MSA) mechanism involving shifted windows [9].

2.3 Relations between our work and existing models. Bandara [10] introduced a multi-layer Siamese encoder empowered by self-attention mechanisms to extract features related to changes in dual phases. This framework incorporates four modules with distinct characteristics to compute feature disparities across multiple scales. Finally, a lightweight MLP decoder is employed to amalgamate previous features and generate CD predictions. Bandara's research findings served as a significant source of inspiration for our study. In contrast, we adopted MSA mechanisms based on shifted windows in the coding stage to avoid the heavy calculation load of global attention during feature extraction. This reduced the calculation complexity of our model. Additionally, it provided favorable conditions for the extraction and output of high-resolution images. Moreover, a cascade feature fusion decoder was introduced in the decoding stage. Compared with the lightweight MLP-based decoder, we improved the interconnection between decoding layers through top-down cascade paths and dense cascades. The feature

classification head was used to process the fused images, thus achieving the semantic feature change output of high-quality and high-resolution images.

3. Proposed Approach

3.1 Architecture. The present research introduces a CD model tailored for high-resolution remote sensing images, utilizing the Siamese network as its overarching architecture. The detailed model configuration is illustrated in Fig. 2. SiamFormer is a standard model based on encoding and decoding architectures, mainly including three components:

·Input($H\times W\times 3$) & Output($H\times W\times 1$)
·Hierarchical Swin Transformer Encoder
·Cascading Feature-Fused Decoder

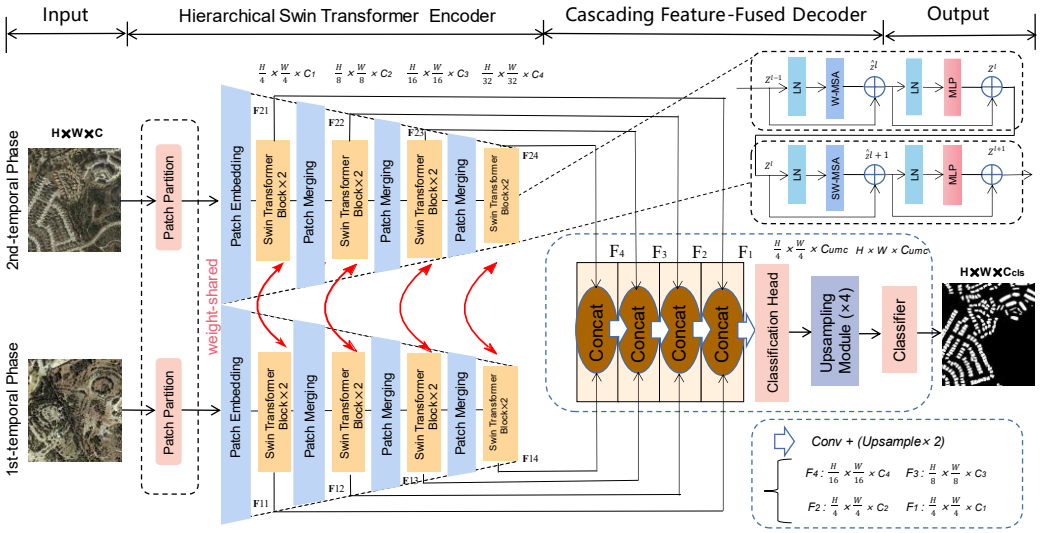

Fig. 2. SiamFormer model structure

As depicted in Fig. 2, the dual-temporal images (1st-tp *T1* and 2nd-tp *T2*) are initially fed into a multi-layer Siamese encoder. This encoder comprises two parallel sub-networks featuring identical structures and shared weights. Each sub-network is composed of four stages, with each stage comprising Patch Merging and Swin Transformer layers. In addition, the convolution-free down-sampling technique is adopted in each coding stage. Compared with the ViT, the most significant difference lies in the construction of hierarchical feature maps. During feature extraction, the feature mapping with hierarchical structures was generated by merging and down-sampling in each layer. Subsequently, a cascaded feature fusion decoder was employed to progressively integrate the multi-scale high-resolution features obtained from the multi-layer Swin Transformer encoder.

Finally, a predicted CD mask code was generated by the classification head and up-sampling.

3.2 Hierarchical Swin Transformer Encoder. As mentioned previously, the model adopts a hierarchical Swin Transformer encoder. For a given dual-phase image, the process can generate high-resolution coarse features and low-resolution fine features akin to those required for convolution in CD tasks. Specifically, the dimension size of the dual-phase image is $H\times W\times 3$. The feature map F_i (i, j={1,2,3,4} and $C_{i,j+1}>C_{i,j}$) with the output size of $(\frac{H}{2^{j+1}} \times \frac{W}{2^{j+1}}) \times C_{ij}$ generated through the hierarchical Transformer encoder was hierarchically fused to the cascade feature fusion decoder for further processing to obtain the change map.

3.2.1 Patch Partition & Patch Embedding. Patch Partition: Patch Partition: The RGB images, sized $H\times W\times C$, are partitioned into non-overlapping, equally-sized $N\times(P^2\times C)$ patches (or 4×4 blocks), where N denotes the number of tokens and P represents the image size. Each $P^2\times C$ patch is treated as a patch token, resulting in a total of N patch tokens (the effective input sequence length of the Transformer). Specifically, in patches with a dimension of P^2 and C channels, the feature dimension size of each patch after flattening is $P\times P\times C$. There are $N=\frac{H}{P}\times \frac{H}{P}$ patch tokens in total. In other words, each image with a size of $H\times W\times C$ is processed into $\frac{H}{P}$ patches. Each patch is flattened to a token vector with $P\times P\times C$ dimensions. Similar to flattened patches in ViT, a flattened 2D patch sequence with $N\times(P^2\times C)$ dimensions can be obtained.

Linear Embedding: The tensor with a dimension size of $N\times(P^2\times C)$ is projected to any dimension (C) so a linear embedding with a dimension size of $(\frac{H}{P} \times \frac{H}{P})\times$C can be obtained.

***3.2.2 Patch Merging*.** In most convolution neural networks, down-sampling through feature mapping is accomplished by convolution. In the Swin Transformer, Patch is the smallest unit in the feature map. So, Patch merging can be used for convolution-free down-sampling. For example, in a 10×10 feature map, there will be $10\times 10 = 100$ patches. In this case, patch merging is performed by the grouping of $n\times n$ adjacent patches and the splicing based on depth. This can successfully promote n times of effective down-sampling for the input. The input is converted from $H\times W\times C$ to $(\frac{H}{n}\times\frac{W}{n})\times(2nC)$, where H, W, and C represent the height, width, and depth of the channel, respectively.

***3.2.3 Swin Transformer Block*.** The conventional MSA mechanism utilized in ViT enables global self-attention computations as illustrated in Fig. 3(b). At the beginning, 16 times downsampling is done through convolution, and then the size of the feature map remains unchanged throughout the entire process. Therefore, the calculation of Attention is performed on the entire feature map, resulting in a large computational load which makes it unsuitable for high-resolution image processing.

To eliminate these issues, the Swin Transformer incorporated a W-MSA approach, adopting multi-scale features, initially downsampling by 4 times, then halving the feature map after each stage, and finally downsampling by 16 times as illustrated in Fig. 3(a). Divide windows within each feature map and perform attention calculation on the elements inside each window instand of calculation the entire feature map, thus greatly reducing computational complexity. However, the W-MSA module is not able to interact with the information due to the segmentation of the window, therefore, the SW-MSA module is proposed as shown in Fig. 3(c).

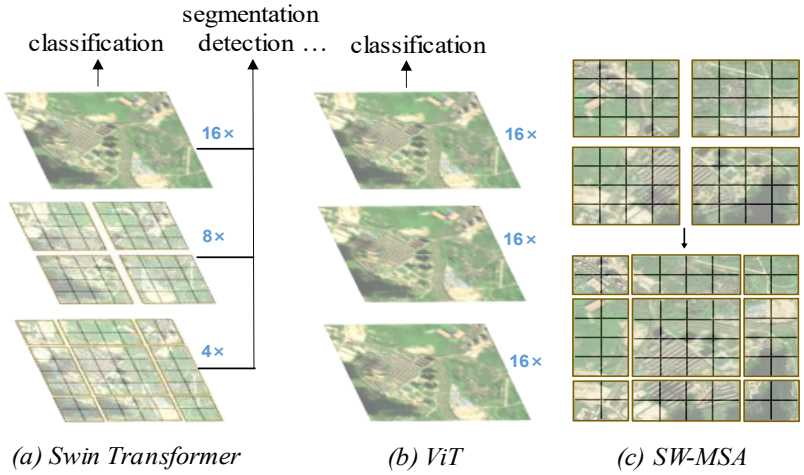

Fig. 3. (a). Based on Swin Tranformer feature processing module (b). Based on ViT feature processing module (c) The shifted window-based multi-head self-attention mechanism

The orange border area represents a window, and the black border area represents a patch as shown in Fig. 3. The input image can be split into non-overlapping windows via the W-MSA mechanism. Based on that, the self-attention of all tokens inside the window can be calculated. In this scheme, the patch represents a subset of these windows. Given that the window size remains constant across the network, the computational complexity of the window-based MSA scales linearly with respect to the number of patches in the image. The calculation burden can be significantly reduced during processing, and it is effective for feature extraction of high-resolution images.

If an image has the dimension size of $H\times W\times C$ and includes $H\times W$ patches, where each window contains $M\times M$ patches, the complexity of MSA and W-MSA can be computed as follows.

Firstly, the operational formula of the MSA module proposed in Transformer can be expressed as:

$$Attention(Q,K,V)=Soft\max\left(\frac{QK^{T}}{\sqrt{d}}\right)V \tag{1}$$

where, $Q, K, V \in R^{M^2 \times d}$ represent the query, key, and value matrices, respectively, M^2 and d denote the number of patches in a window and the dimension size of the query or key.

Subsequently, the computational complexity can be determined using the subsequent attention formula:

$$Q = xW^Q, K = xW^K, V = xW^V \Omega(1) = 3HWC^2 \tag{2}$$

$$QK^T \qquad \Omega(2) = (HW)^2 C \tag{3}$$

$$Z = QK^T V \qquad \Omega(3) = (HW)^2 C \tag{4}$$

$$ZW^Z \qquad \Omega(4) = HWC^2 \tag{5}$$

$$QK^T \times count(Windows) \qquad \Omega(5) = \frac{H}{M}\frac{W}{M}(M^2)^2 C \tag{6}$$

After the integration of the above results, new formulas are obtained:

$$\Omega(MSA) = 4HWC^2 + 2(HW^2)C \tag{7}$$

$$\Omega(W-MSA) = 4HWC^2 + 2\frac{H}{M}\frac{W}{M}(M^2)^2 C = 4HWC^2 + 2M^2 HWC \tag{8}$$

Although the calculation amount based on Windows can be reduced, the visual field of each window is restricted. Only the token inside the current window can be exhibited, but the global information cannot be presented. Moreover, the information cannot be exchanged between windows. Based on that, we increased the receptive field and strengthened the interaction between windows through hierarchical structures and SW-MSA mechanisms.

Fig. 4 depicts a pair of consecutive Swin Transformer blocks. Each block encompasses a normalization layer (LayerNorm), an MSA module, a residual connection, and a two-layer MLP featuring GELU nonlinearity. These two consecutive Transformer blocks incorporate both a W-MSA module and an SW-MSA module.

Based on the design mechanism of this sliding partition window, the consecutive Swin Transformer blocks can be represented as follows:

$$\hat{z}^l = W - WSA(LN(z^{l-1})) + z^{l-1} \tag{9}$$

$$z^l = MLP\left(LN\left(\hat{z}^l\right)\right) + \hat{z}^l \tag{10}$$

$$\hat{z}^{l+1} = SW - MSA(LN(z^l)) + z^l \tag{11}$$

$$z^{l+1} = MLP\left(LN\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1} \tag{12}$$

Here, $\hat{z}^l$ and z^l denote the output features of the (S) W-MSA module and the MLP module of the l^{th} block, respectively.

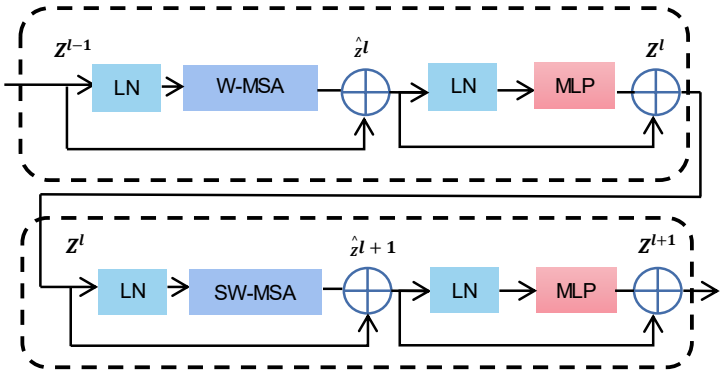

Fig. 4. Swin Transformer blocks

3.3 Cascade Feature-Fusion Decode. The cascade feature fusion decoder, as suggested by SiamFormer, was utilized for extracting and integrating multi-scale features obtained from the hierarchical Swin Transformer encoder, so as to gradually enhance the salience of the change region and predict the CD image. Eventually, classification and dimension reduction were achieved through the convolution operation by the multi-scale classification head and up-sampling module. Based on that, a change map was generated.

4. Experimental results and analysis

4.1 Datasets and Preprocessing. Our study utilized two well-regarded datasets in the domain of remote sensing image CD, namely LEVIR-CD and CDD.

The CDD dataset, sourced from Google Earth, features three distinct categories of images. This diversity in change scenarios enables a thorough validation of the proposed CD algorithm's efficiency. The images are standardized to a resolution of 256×256 pixels, comprising a dataset of 10,000 training pairs, 3,000 testing pairs, and 3,000 validation pairs. The spatial resolution is 3-100cm/px, which ensures a high adaptation between the algorithm and hardware.

Differently, the LEVIR-CD database is established based on the Google Earth API. The dual phases of the data span from 5 to 14 years and reflect the maximum land use change consisting of numerous land change maps. The dataset includes 637 pairs of databanks, and the spatial resolution is 50cm/px. In LEVIR-CD, the standard image size is 1*024×1024*px, and there are more than 31,000 independent mark change instances.

4.2 Implementation Details.We used Pytorch and NVIDIA GeForce RTX 3090 24GB VRAM for model training during our experiments. Data enhancement and weighted use of various loss functions were performed during model training,

and favorable results were obtained. Besides, the Adam gradient-based momentum optimizer was used to perform model training, and the weight attenuation was 0.0001. Additionally, our approach incorporates the application of the cosine annealing attenuation technique for modulating the learning rate throughout the training phase of the model. After the learning rate was attenuated from 0.0001 (the initial value) to 1, it returned to the original level through readjustment. Subsequently, the current local optimal solution was temporarily excluded to identify the global optimal solution, with 20 rounds as an iteration until the model reached 120 rounds of training. The model training was implemented based on a batch size of 20.

4.3 Performance Metrics. The task of detecting changes in high-resolution remote sensing imagery can be equated to performing binary classification on pixels. In the analysis of this binary classification, we followed Bandara's work that employ evaluation metrics such as precision (p), recall (r), F1-score (F1), and the Intersection over Union (IoU). Additionally, we report the number of parameters (Params.), and the floating-point operations (FLOPs).

4.4 Results and Discussion. In this section, we evaluate and compare eight leading-edge models in the domain of CD within remote sensing imagery, focusing on their effectiveness in executing CD tasks.

4.4.1 Comparative Analysis. Table 1 displays a comparison of the outcomes from eight cutting-edge networks specialized in CD on the testing subsets from LEVIR-CD and CDD change detection datasets. By comparing data regarding performance metrics, we concluded that the network introduced in this research outperformed competing models with respect to F_1, IoU, and Rec. In addition, compared to some other Transformer-based models, such BIT and ChangeFormer, our proposed model basically leads in every metric. On the LEVIR-CD dataset, our model achieves the leading F1 value, which is higher than the second-best ChangeFormer by 0.52 and third-best SCADNet by 0.60. In particular, we achieve first place in all metrics on the CDD dataset, surpassing the second-best model by 0.30 in F1, 0.7 in Pre., 1.42 in Rec. and 2.48 in IoU.

Furthermore, we also concluded that during the resolution of more complex detection problems, such models could exhibit inferior capabilities in feature extraction, leading to diminished accuracy and lackluster robustness. Among them, in our comparative experiment, fewer feature extraction layers were adopted in FC-EF, and the number of channels in the last layer was small, resulting also in smaller models. Compared with similar networks based on a pure Transformer coder, our network not only improves on the recognition metrics, but also in the number of Params. was less than half of those networks. Especially, the computational complexity of our network on FLOPs was far lower than that of other models. This result indicated that the calculation strategy of our model was indeed effective,

which can significantly reduce the calculation complexity and parameters and contribute to optimal CD performance.

Table 1

Model Performance Compared on LEVIR-CD and CDD Dataset

Method	*LEVIR-CD*				*CDD*				*Params.*	*FLOPs*
	Pre./%	*Rec./%*	*F1/%*	*IoU/%*	*Pre./%*	*Rec./%*	*F1/%*	*IoU/%*	*(M)*	*(G)*
FC-EF[1]	80.24	70.31	74.95	59.93	66.73	54.08	59.74	42.59	1.35	1.78
STANet[11]	91.90	85.00	88.10	79.12	92.28	85.44	88.61	80.12	16.93	6.58
IFNet[12]	94.02	82.93	88.13	78.77	94.96	86.08	90.30	-	50.71	41.18
SNUNet[13]	89.18	87.17	88.16	78.83	95.60	94.90	95.30	-	12.03	27.44
BIT[14]	89.24	89.37	89.31	80.68	88.97	82.73	85.74	75.03	3.55	4.35
SCADNet[15]	90.14	91.74	90.32	90.56	-	-	-	-	66.94	70.72
ChangeFormer[16]	92.05	88.80	90.40	82.48	94.50	93.52	94.23	89.09	41.00	101.4
SiamFormer	91.45	90.39	90.92	83.35	96.30	94.90	95.60	91.57	27.70	7.21

4.4.2 Comparison on the LEVIR-CD and CDD Dataset. As depicted in Fig. 5, 2,048 pairs of LEVIR-CD datasets and 2,999 pairs of CDD datasets were tested in model training. Moreover, they were also compared separately according to the five groups of comparative references we established during the experiments. Our comparative experiment showed that SiamFormer performs best in change map prediction, far exceeding the second model, with a high similarity with the true value. Compared with the comparative model, there is almost no noise in our model, and the size of the detected object is close to the true value.

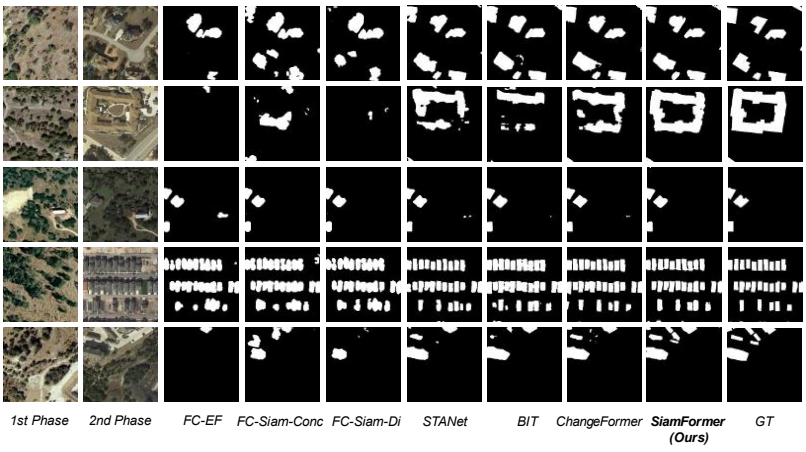

Fig. 5. Comparative performance analysis of various models on the LEVIR-CD dataset for change detection tasks.

Fig. 6 shows a comparative prediction chart based on CDD datasets, confirming that our model presents favorable results. Especially in the boundary of

the image, our model is almost the same as the true value in texture and size, which outperforms the experimental results of other comparative models.

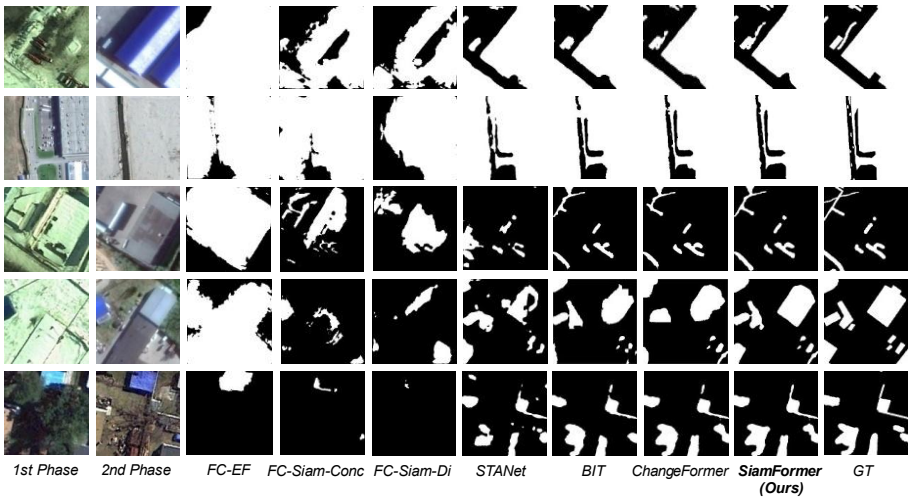

Fig. 6. Comparative performance analysis of various models on the CDD dataset for change detection tasks.

5. Conclusion

In this paper, we introduced the SiamFormer, a dual-phase Siamese network framework based on the Swin Transformer, to improve the resolution of change detection in large datasets and reduce model complexity. By combining the pre-training model of the Swin Transformer with the Siamese network, we developed a multi-layer encoder to extract high-resolution features of dual phases. Cascade feature fusion coding was introduced for effective classification of change regions between images. Experimental results indicated that the Siamese network's layered coding and Swin Transformer's feature fusion decoding capabilities accurately identified change targets without deep CNNs. This approach significantly reduces model complexity. Future research should focus on exploring weakly supervised or unsupervised learning for model training to further reduce the reliance on massive data for high-resolution change detection.

Acknowledgments

This work was financially supported by the Science and Technology Development Project of Henan Province of China (202102110115 and 242102320301) and the Science and Technology Development Project of Anyang of China (2023C02GH061 and 2023C02GH062).

REFERENCES

[1]. *Caye Daudt R, Le Saux B, Boulch A.* Fully convolutional Siamese networks for change detection//2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, 2018:4063- 4067.

[2]. *Ronneberger O, Fischer P, Brox T*. U-net: Convolutional networks for biomedical image segmentation//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015:234- 241.

[3]. *Peng D F, Zhang Y J, Guan H Y*. End-to-end change detection for high resolution satellite images using improved UNet ++ . Remote Sensing, 2019,11(11):1382.

[4]. *Chen H, Shi Z W*. A spatial- temporal attention- based method and a new dataset for remote sensing image change detection. Remote Sensing, 2020,12(10):1662.

[5]. *Dosovitskiy A, Beyer L, Kolesnikov A, et al.* An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[6]. *Liu Z, Lin Y, Cao Y, et al.* Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

[7]. *Xie E, Wang W, Yu Z, et al.* SegFormer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems, 2021, 34: 12077-12090.

[8]. *Vaswani A, Shazeer N, Parmar N, et al.* Attention is all you need. Advances in neural information processing systems, 2017, 30.

[9]. *Cao H, Wang Y, Chen J, et al.* Swin-unet: Unet-like pure transformer for medical image segmentation[C]//Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III. Cham: Springer Nature Switzerland, 2023: 205-218.

[10]. *Bandara W G C, Patel V M*. A transformer-based siamese network for change detection[C]//IGARSS 2022 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2022: 207-210.

[11]. *Zhu X, Su W, Lu L, et al.* Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159, 2020.

[12]. *Carion N, Massa F, Synnaeve G, et al.* End-to-end object detection with transformers[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 213-229.

[13]. *Zhang C, Yue P, Tapete D, et al.* A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 166: 183-200.

[14]. *Chen H, Qi Z, Shi Z.* Remote sensing image change detection with transformers. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-14.

[15]. *Xu C, Ye Z, Mei L, et al.* SCAD: A Siamese cross-attention discrimination network for bitemporal building change detection. Remote Sensing, 2022, 14(24): 6213.

[16]. *Bandara W G C, Patel V M.* A transformer-based siamese network for change detection[C]//IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2022: 207-210.