Global Context-Enhanced Network for Pixel-Level Change Detection in Remote Sensing Images

Zixue Zhao, Zhengpeng Li, Member, IAENG, Jiawei Miao, Kunyang Wu, and Jiansheng Wu*

Abstract—**Despite the ongoing advancements in deep learning, challenges persist in the domain of change detection in remote sensing imagery. Objects with intricate structures and features may exhibit different shapes or appearances at different times or spatial locations. While most models aim to improve the performance of change detection tasks, these enhancements may lead to significantly increased computational efficiency. In this paper, we propose a global context enhancement network. Firstly, we use ResNet18 to extract dual-temporal features, which are then represented as concise semantic labels by an image semantic extractor. Subsequently, we process these semantic labels through a contextual transformer encoder to generate more refined remote sensing semantic labels enriched with abundant contextual information. The refined semantic labels are integrated with the original features and processed through a Transformer decoder to generate enhanced dual-temporal feature maps. Finally, through the processing of the classification head, we obtain pixel-level predictive images. Extensive experiments conducted on two public change detection datasets yielded impressive results, achieving an F1 score of 89.95% on the WHU-CD dataset and 95.16% on the SVCD dataset. When compared to state-of-the-art change detection models, our approach not only achieves significant performance gains but also maintains relatively high computational efficiency. Our method excels in capturing relevant features and relationships within input data, thereby enhancing the model's ability to represent relationships between different features. This results in a significant performance improvement without adding to the computational complexity.**

Index Terms— **Change Detection, Global Context Information, Attention Mechanism, Computational Efficiency, High-Resolution Remote Sensing Images.**

I. INTRODUCTION

Remote sensing image change detection [1] is a technique used to identify and analyze real-world object variations by comparing remote sensing images captured at different points in time. In the field of remote sensing imagery, change detection has garnered widespread attention due to its significance in understanding the evolution and alterations within geographical regions. Research in this field has played a crucial role in practical applications, including land change detection [2, 3], environmental change monitoring [4], disaster assessment [5], urban change studies [6], and ecological environment research [7]. While manual annotation is typically required for most high-resolution remote sensing images, the development of deep learning technologies has increasingly highlighted methods for change detection.

The deep learning-based change detection methods still face a series of challenges, mainly covering the following two aspects: 1) objects present in some remote sensing image scenes exhibit complex structures and features; 2) the same object may appear with different shapes or appearances at different times or spatial locations. Due to variations in lighting angles, buildings in the reference image and target image may have different color features.

In recent times, a series of deep convolutional neural networks (CNNs)[8-10] have been applied to high-resolution remote sensing image change detection. However, these CNNs still face certain limitations, preventing them from fully addressing the two challenges mentioned earlier. Some models attempt to enhance feature extraction by either stacking more convolutional layers or utilizing dilated convolutions [11] to increase the receptive field. Meanwhile, other models introduce attention mechanisms such as spatial attention, cross-entropy attention, and channel attention [12, 13] to broaden the receptive field, thus preserving global contextual information. Nevertheless, many existing methods still encounter difficulties when handling high-resolution remote sensing images. They either focus solely on enhancing the internal features of each temporal image, neglecting the correlations between global contextual information, or emphasize only on weight-fusing dual temporal features through attention in either the channel or spatial dimensions to enhance the interrelationships of contextual information. Consequently, existing methods often struggle to effectively correlate spatiotemporal feature information, especially when dealing with remote sensing images exhibiting complex changes. This issue becomes particularly evident.

To address the limitations and challenges outlined, we propose the Global Context Enhancement Net (GCENet) model, designed to efficiently correlate global contextual features in dual-temporal images. Leveraging a deep convolutional neural network to extract semantic features from input images, advanced semantic features are transformed into a set of semantic labels through max-pooling. The model enhances the fea-

Manuscript received March 14, 2024, revised July 11, 2024.

Zixue Zhao is a graduate student of University of Science and Technology Liaoning, Anshan, 114051, China. (e-mail: mszzx1998@163.com).

Zhengpeng Li is a doctoral student of University of Science and Technology Liaoning, Anshan, 114051, China. (e-mail: lkdlzp0901@163.com).

Jiawei Miao is a doctoral student of University of Science and Technology Liaoning, Anshan, 114051, China. (e-mail: 455605116@qq.com)

Kunyang Wu is a doctoral student of Jilin University, Changchun, China. (e-mail: wuky23@mails.jlu.edu.cn)

Jiansheng Wu is a professor of University of Science and Technology Liaoning, Anshan, 114051, China. (corresponding author, e-mail: ssewu@163.com).

ture representation of the original pixel space by exploiting the relationship between pixels and semantic labels. Subsequently, we introduce the Contextual Transformer (COTR) encoder, utilized for contextual modeling of semantic labels, to generate enriched semantic labels with contextual information. These are then reprojected back into pixel space through a Transformer decoder, enhancing both feature information extraction and contextual correlation. Finally, a prediction module is employed for change prediction. The COTR encoder effectively prevents a reduction in the correlation between feature channels, thereby strengthening the capture of sufficient feature information.

The main contributions of this paper are summarized as follows:

1) We propose a novel framework for remote sensing change detection, termed the GCENet. GCENet adopts a Transformer decoder structure to map semantic labels containing rich global information back to pixel space, thereby enhancing the original pixel features. This approach efficiently and effectively addresses the task of change detection.

2) We designed an image semantic extractor to process input images of the same location at different times, generating a set of concise semantic labels. This allows for the effective capture of semantic information at the same location across different time points.

3) We establish a COTR encoder structure and propose a context feature interaction attention mechanism. Introducing convolutional operations into the attention mechanism enables more effective capture of relevant features and relationships within the input data. This contributes to a better understanding of contextual relationships among features, allowing the model to adapt more effectively to complex input data.

Numerous experiments were conducted on the WHU-CD and SVCD datasets, comparing our method with other change detection models. The results confirm the effectiveness and efficiency of our approach. Additionally, our model outperforms other change detection models in terms of parameter count and floating-point operation count, demonstrating superior performance.

II. Related Work

A. Convolution-Based Models for Remote Sensing Image Change Detection

With the evolution of deep learning, an increasing number of methods have been applied to remote sensing image change detection. To the best of our knowledge, the first work to apply convolutional neural networks for change detection tasks was proposed by Caye Daudt et al.[14]. In this work, the authors proposed two distinct approaches. The first method employed a pixel-level fusion strategy, FC-EF, which merged images from two different time points into a single input processed through a U-Net. The second method adopted a feature-level fusion strategy, utilizing twin U-Nets to process each image separately, extracting multi-level features that were subsequently fused through concatenation and subtraction. Lei et al.[15] introduced a pyramid pooling fully convolutional network, efficiently explore the surrounding environment of dual-temporal remote sensing images through multiple convolutions. This achieved a well-balanced trade-off between enlarging the perception range and fully utilizing contextual information. Tang et al.[16]. proposed an unsupervised change detection method based on graph convolutional networks and metric learning, which generates reliable difference maps through Siamese Fully Convolutional Network (FCN), multi-scale dynamic graph convolutional networks and metric learning. Wu et al.[17] proposed an unsupervised change detection method, which extracts features from multi-temporal ultra-high resolution images through deep Siamese networks, realizes unsupervised binary and multi-category change detection, leading to effective and robust experimental results. Shen et al.[18] proposed an unsupervised change detection method based on an improved non-subsampled Shearlet transform and a multi-scale feature fusion convolutional neural network. At the same time, a multi-scale feature fusion block was designed to retain more detailed information, aiming to achieve the goal of obtaining change detection results directly from the original image.

B. Attention Mechanism-Based Models for Remote Sensing Image Change Detection

In recent years, attention mechanisms have demonstrated significant effectiveness in capturing crucial differences in feature space and channels across various computer vision tasks. Notably, self-attention, spatial attention, and other attention mechanisms enable the modeling of global spatial relationships, effectively assisting networks in recognizing objects undergoing change and those remaining unchanged. Zhang et al.[19] introduced channel and spatial attention mechanisms and applied them at various levels of the decoder. By weighting the fusion of multi-layer deep features from the original image, the network was able to focus more on capturing key information in changing regions. Similarly, Liu et al.[12] also used the dual attention module. Jiang et al.[13] introduced a global co-attention mechanism to effectively improve the long-range dependence of features, and then further enhanced the model's feature extraction of target information by training convolutional neural networks in a pyramid structure to capture possible changes. Chen et al.[20] introduced a time attention mechanism. The time attention mechanism involves extracting features from dual-temporal feature maps, generating a query matrix and a value matrix, and performing dot product operations. The resulting numerical matrix was then matched with the feature map generated from the target image. Lei et al.[21] proposed a network for high-resolution remote sensing image change detection. This network achieved more accurate detection results and reduced computational complexity through difference enhancement, spatial-spectral non-locality, and asymmetric dual convolution combined with spatial multi-scale features.

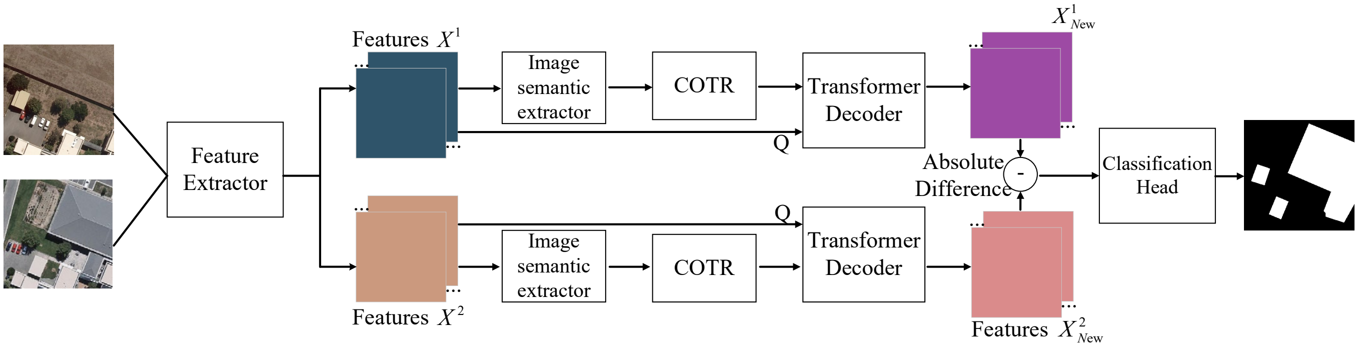


Fig. 1. Diagram of the GCENet Model.

C. Transformer-Based Models for Remote Sensing Image Change Detection

With the introduction of the Transformer [22, 23], it has not only found widespread application in natural language processing but has also exhibited superior performance in the visual domain compared to models based on convolution, as observed in recent years. Chen et al.[24] extracted bi-temporal features using ResNet, converted them into semantic tags, and then enhanced the features with global information through the Transformer structure for change detection. Bandara et al.[25] initially utilized the Transformer for feature extraction, followed by the feature difference module to calculate feature discrepancies at different scales and generate prediction results. By leveraging the Swin Transformer [26] and a Siamese structure, Liu et al.[27] achieved parallel processing of dual-temporal images and extraction of multi-scale features. Dai et al.[28] introduced the Vision Transformer into change detection, used multi-scale feature differentiation to improve the discrimination of multi-level context information, resulting in high-precision change detection tasks being accomplished at a faster speed. Feng et al.[29] proposed a change detection network that combines Transformer and convolution. This network introduced an intra-scale cross-interaction mechanism to effectively capture both local and global features.

III. Efficient Change Detection Based on Global Context Encoder

The overall flow of our model for GCENet is shown in Fig. 1. We aim to incorporate the advantages of convolution and transformer into the change detection model. Our model consists of five main components:

1) Feature extractor: Utilizing ResNet18 as the backbone, it is responsible for extracting the bi-temporal feature maps.

2) Image semantic extractor, which takes images of the same location at different times as input and produces a concise set of semantic labels.

3) COTR encoder. Performs context modeling for the semantic labels to generate semantic labels with rich global information.

4) Transformer decoder, maps semantic labels with rich global information back to the pixel space and enhances the original pixel features.

5) Classification header, which generates pixel-level change predictions to identify whether a change has occurred at each pixel point in the image.

The process of running the model begins with the feature extractor module, which performs feature extraction for each input image. Subsequently, these features are fed into the model, which undergoes an image semantic extractor to generate a concise set of semantic labels. Next, we improve the new raw pixel features using the COTR encoder and the Transformer decoder. Finally, pixel-level features are predicted by a classification header consisting of convolution, normalization, and activation functions to generate pixel-level change predictions.

A. Image Semantic Extractor

To describe the changes in interest in the images, we employ a set of high-level concepts, namely semantic labels. To achieve this objective, we introduce the Image Semantic Extractor, as depicted in Fig. 2. Its task is to extract concise semantic labels from the feature maps at each time step. Differing from tokenizers in natural language processing, this extractor does not segment the input image into visual words and generate corresponding token vectors for each word. Instead, it aggregates the feature maps in the spatial dimension by learning a set of spatial attention maps. This process generates a set of features corresponding to the semantic labels of the image. These generated semantic labels can be shared across images at different time steps, providing more detailed and global information for the change detection task.

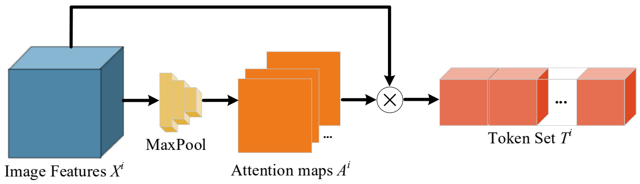


Fig. 2. Image semantic extractor.

We define the dual-temporal feature maps as $\{X^1, X^2\} \in \mathbb{R}^{HW\times C}$, where H, W, and C represent the height, width, and channel dimensions of the feature maps, respectively. Simultaneously, we introduce two sets of labels $\{T^1, T^2\} \in \mathbb{R}^{HW\times C}$.

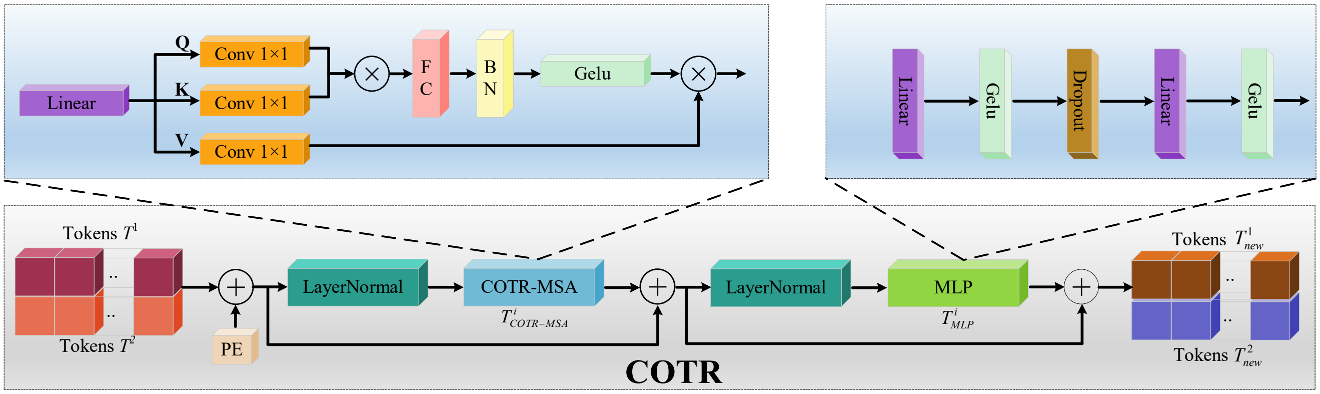


Fig. 3. Contextual Transformer Encoder.

When processing each pixel X_p^i on the feature map $X^i(i \in 1,2)$, we use an adaptive max-pooling operation to create semantic groups, where each group represents a semantic concept. Subsequently, we utilize the softmax function to operate on the HW dimensions of each semantic group, calculating an attention map. Finally, we use the attention map to compute the weighted average of pixels to obtain a concise label set $T^i(i \in 1,2)$. The formula is as follows:

$$T^i = (A^i)^T X^i \tag{1}$$

$$A^i = \sigma(Maxpool(X^i)) \tag{2}$$

where $Maxpool(\cdot)$ represents the adaptive max-pooling operation, and $\sigma(\cdot)$ denotes the softmax function. The softmax function is used to normalize each semantic group and generate an attention map $A^i \in \mathbb{R}^{HW \times C}$. Finally, the label set T^i is obtained by multiplying the attention map A^i with the input feature map X^i.

B. Contextual Transformer Encoder

As illustrated in Fig. 3, we propose using the COTR to capture the contextual relationships among these semantic labels. In a token-based spatiotemporal space, we believe that utilizing the COTR encoder can effectively model global semantic relationships, thus producing token representations with abundant contextual information for each time step. The COTR encoder module is primarily composed of *n* layers of COTR-MSA, *n* layers of MLP modules, and a normalization layer.

We start by merging these two sets of labels T^1 and T^2, to form a unified label set $T \in \mathbb{R}^{HW \times C}$. Subsequently, we introduce the COTR encoder module, applying layer normalization before the COTR-MSA and MLP modules. Pre-normalization has been proven to be advantageous in improving model stability and performance. Additionally, we use a residual connection to add the outputs of these two modules with the feature map before normalization, generating a completely new label set T_{new}. Finally, we partition these labels into two subsets $T_{new}^i(i \in 1,2)$. This series of processing steps contributes to enhancing model performance by enabling the output features to balance local and global contextual information representations, thereby enriching the semantic information. The computational formula is as follows:

$$T^i_{COTR-MSA} = MSA_{COTR}(LN(T^i)) \tag{3}$$

$$T^i_{MLP} = MLP(LN((T^i_{COTR-MSA} + T^i)) \tag{4}$$

$$T^i_{new} = T^i_{COTR-MSA} + T^i_{MLP} \tag{5}$$

In the classic Transformer architecture, the multi-head self-attention mechanism learns weight matrices for query vector Q, key vector K, and value vector V by providing multiple representation subspaces. It generates self-attention scores by computing the dot product of the query vector Q and the key vector K across the entire sequence, thereby assessing the correlation between input features. This self-attention mechanism is employed to model the correlation among all tokens, allowing the model to effectively integrate global information and inter-channel information to search for information across global space and channels. This ensures that the model focuses on contextual information so that the final output features encompass global information. Finally, the SoftMax function is applied to prevent gradient vanishing, and the result is multiplied by the value vector V to obtain the ultimate output.

We observed that in the classic multi-head self-attention mechanism, the direct dot product operation between Q and K processed by linear layers may reduce the correlation between feature channels. Additionally, each self-attention head focuses on only a subset of input tokens, which could impact network performance in certain cases, especially when the channel dimensions of each subset are relatively low. This situation can lead to insufficient information capture in the dot product between Q and K, thereby diminishing the ability to gather contextual information. Specifically, for data like high-resolution remote sensing images, increasing spatial or channel dimensions may result in a rapid rise in computational complexity. Furthermore, the handling of V in the classic self-attention mechanism also has limitations. V is typically used for a linear combination and weighting of the representations of each token. However, direct linear combinations may not fully explore complex relationships in the input data, thus limiting the model's understanding of certain tasks or data. In our ablation experiments, we further investigated the impact of linear layer processing and dot product operations in the classic multi-head self-attention mechanism on model performance. The specific effects of these changes on model

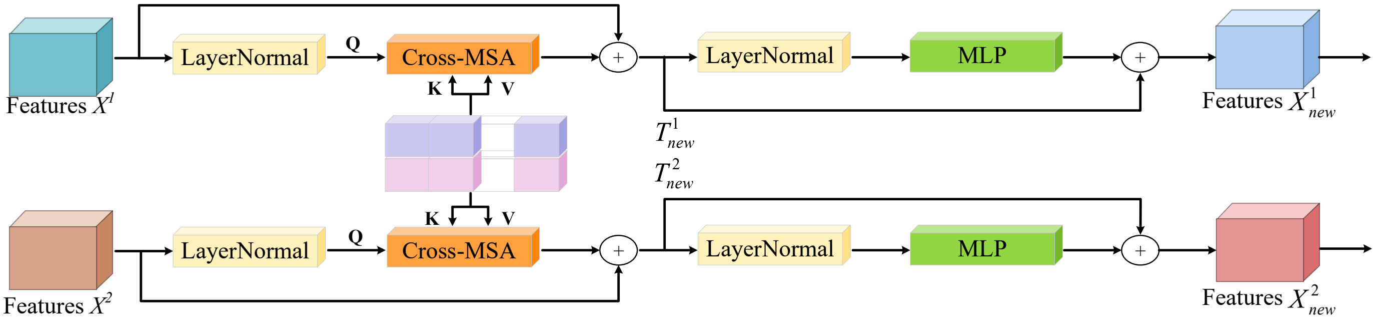


Fig. 4. Transformer decoder.

performance and contextual information retrieval are detailed in Table IV based on our experimental results.

To address these challenges, we introduced the COTR-MSA mechanism. The COTR-MSA executes multiple independent attention heads in parallel, concatenates their outputs, and then performs projection. The advantage of COTR-MSA lies in its ability to simultaneously focus on information from different representation subspaces at different positions. By introducing convolution operations on Q, K, and V, the model better captures relevant features and correlations in the input data. The introduction of these convolution operations is significant in our experiments, as detailed in Tables I and II. This enhances the model's ability to model relationships between different features, further strengthening the correlation between feature channels, providing the model with a deeper understanding of the data. The formula is as follows:

$$\begin{aligned} MSA_{COTR}(T^{(l-1)}) &= Concat(head_1, head_2, ...head_h)W^O, \\ head &= Att_{CFI}(T^{(l-1)}W^Q, T^{(l-1)}W^K, T^{(l-1)}W^V) \end{aligned} \quad (6)$$

where $W^Q, W^K, W^V \in \mathbb{R}^{C \times d}$, $W^O \in \mathbb{R}^{hd \times C}$ is the weight matrix.

Firstly, we map the input token sets T^1 and T^2 to $Q, K, V \in \mathbb{R}^{b \times h \times n \times d}$ through a linear layer, where b represents the batch size, h denotes the number of attention heads, n is the sequence length of the token set, and d is the feature dimension. Next, pointwise convolution operations are applied separately to Q, K, and the numerical values V through the COTR-MSA. The purpose of this step is to introduce local perception capability, enabling the aggregation of semantic information across channels for different representations within the input data. A further extension to the pointwise product of Q and K is performed through a fully connected layer to ensure no reduction in dimensionality. Finally, the results of the dot product are optimized through batch normalization and softmax operations to ensure they effectively match the feature information within. The entire process aims to thoroughly explore the intricate relationships in the input data. For this purpose, we propose a novel attention mechanism called Contextual Feature Interaction (CFI) Attention.

$$Att_{CFI}(Q, K, V) = \sigma(BN(FC(\frac{conv(Q)conv(K)^T}{\sqrt{d}})))conv(V) \quad (7)$$

The MLP module primarily consists of a linear layer, GELU activation function, and a dropout layer. The dimensions of both the input and output are C, while the internal dimension of the linear layer is 2C. The formula for the MLP module is as follows:

$$MLP(T^{(l-1)}) = Dropout(Dropout(GELU(T^{(l-1)}W_1))W_2) \quad (8)$$

where $W^1 \in \mathbb{R}^{C \times 2C}$, $W^2 \in \mathbb{R}^{2C \times C}$ represents the weight matrix.

C. Transformer Decoder

As shown in Fig. 4, we introduce a Siamese Transformer decoder structure to map the compact high-dimensional semantic information in the mark set $T_{new}^i (i \in 1,2)$ to a low-dimensional pixel space, thereby obtaining a pixel-level feature representation in remote sensing images. The dual-temporal feature maps X^1 and X^2 are separately input into a Siamese Transformer decoder. By exploiting the relationship between each pixel feature and the mark set, we ultimately obtain more refined fine-grained features X_{new}^1 and X_{new}^2. The Transformer decoder we use consists of m layers of Multi-head Cross-Attention, m layers of MLP modules, and LN layers. In the Cross-Attention mechanism, the Q comes from the original image pixels, while K and V come from $T_{new}^i (i \in 1,2)$. The MLP modules and LN layers used are the same as those in the COTR encoder. For each layer l, the formula for Multi-head Cross-Attention is as follows:

$$\begin{aligned} MA(X^{i,(l-1)}, T_{new}^i) &= Concat(head_1, head_2, ...head_h)W^O \\ head &= Att_{CFI}(X^{i,(l-1)}W^Q, T_{new}^i W^K, T_{new}^i W^V) \end{aligned} \quad (9)$$

D. Other Details

CNN Backbone Network: We employed an enhanced ResNet18 as the backbone of the CNN to extract temporal-cross-dimensional feature maps. The standard ResNet18 comprises four stages, each undergoing downsampling twice. We modified the downsampling stride to 1 for the last two stages and added a pointwise convolutional layer after ResNet. Subsequently, a bilinear interpolation layer was applied to generate the output feature map, mitigating the loss of spatial details.

COTR encoder and Transformer decoder: Based on the parameter ablation experiments in Table V, we set the number of layers for the COTR encoder (n) and Transformer decoder (m) to 1 and 2, respectively.

Classification Head: High-level refined features X^1, X^2, extracted through the CNN backbone network, COTR encoder, and Transformer decoder, undergo upsampling to obtain the original images $X_*^1, X_*^2 \in \mathbb{R}^{H_0 \times W_0 \times C}$, where H_0 and W_0 represent the height and width of the original images. Subsequently, pixel-level offset calculation is performed on the two feature maps, followed by the generation of a predicted change probability map $P \in \mathbb{R}^{H_0 \times W_0 \times 2}$ using a change classifier and a non-linear transformation function. The calculation formula is as follows:

$$P = \rho(D(\left|X_*^1 - X_*^2\right|)) \tag{10}$$

where $\rho(\cdot)$ denotes the sigmoid function and $D(\cdot)$ denotes the change classifier, which consists of a convolutional layer, a BN layer and a RELU activation function.

The loss function, which we use to minimize the cross-entropy loss during training to optimize the network parameters, is given in the following formula:

$$Loss = \frac{1}{H_0 \times W_0} \sum_{h=1,w=1}^{H,W} l(P_{hw}, Y_{hw}) \tag{11}$$

where $l(P_{hw}, y) = -\log(P_{hwy})$ is the cross-entropy loss function and Y_{hw} is the label of the pixel at the coordinate (h, w).

IV. Experimental Results And Analysis

A. Experimental Environment and Parameters

The experiments were conducted on GPU A100. The network parameters are set as follows: we use a stochastic gradient descent (SGD) optimizer with momentum to optimize the model. The momentum of the SGD optimizer is set to 0.9, and the weight decay of the regularization term is set to 5e-4. The initial learning rate is a linear decay of the learning rate, which can be used to gradually reduce the learning rate during the training of the neural network. Rate in order to converge more stably in the later stages of training. The formula is as follows:

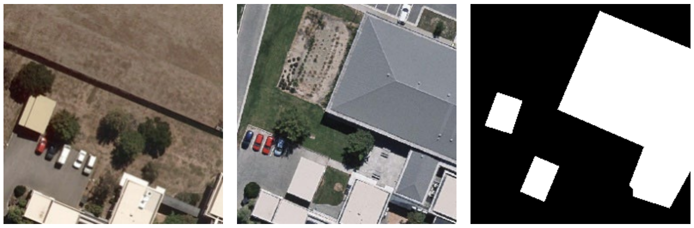
$$lr = 1 - \frac{epoch}{epoch_{\max} + 1} \tag{12}$$

where $epoch$ is the current number of training rounds and $epoch_{\max}$ denotes the total number of training rounds 200, after each training, the best model from the validation set is used to evaluate the test set.

B. Experimental Datasets

We conducted experiments on two change detection datasets:

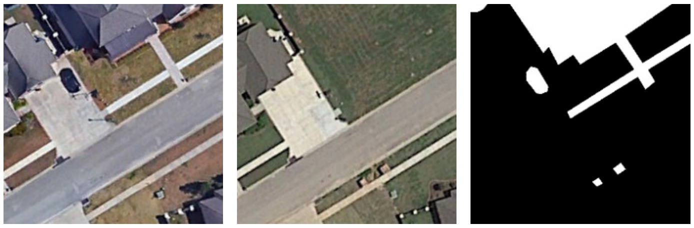
The Wuhan University Building Change Detection (WHU-CD) dataset is designed for remote sensing image change detection, created and distributed by a research team at the School of Remote Sensing Information Engineering of Wuhan University. The dataset consists of a pair of high-resolution aerial images with a size of 15354 × 32507. The focus is on architectural changes. The original images were cropped into 7434 images of size 256 × 256 without overlapping regions. These images were randomly divided into 5947 pairs for the training set, 743 pairs for the validation set, and 744 pairs for the test set. As shown in Fig. 5, the reference image represents the initial image in the dataset, i.e., the image from the first time period of the region. The target image represents the image acquired after the reference image in the dataset.



reference image target image label

Fig. 5. Visualization of WHU-CD dataset.

The Season-Varying Change Detection (SVCD) involves analyzing real remote sensing images obtained from Google Earth that exhibit seasonal changes. The spatial resolution of these images ranges from 3cm to 100cm per pixel. The dataset has 16000 pairs of remote sensing images, each of size 256×256. These pairs are divided into a training set of 10000 pairs, a validation set of 3000 pairs, and a test set of 3000 pairs. Each pair of images contains at least one element of change. The dataset will be labeled as change in the label only if there is a change in the semantic category. Seasonal changes such as vegetation growth, snow cover, and foliage fading etc. will not be labeled as change. This places greater demands on the accuracy of change detection methods and the utilization of semantic information. As shown in Fig. 6..



reference image target image label

Fig. 6. Visualization of SVCD dataset.

C. Evaluation Metrics

We used the F1-score as the main evaluation metric and also used Precision, Recall, Intersection over Union (IoU), and Overall Accuracy (OA). Precision indicates the proportion of correctly predicted positive categories out of all samples predicted by the model to be positive categories. Recall, on the other hand, indicates the proportion of all actual positive category samples that are successfully detected by the model as positive category samples. The F1 value, on the other hand, is the harmonized average of the precision and recall. The cross-merge ratio is used for precision in the change detection task, while the overall accuracy is used for overall performance evaluation.

$$F1 = \frac{2(Precision \cdot Recall)}{Precision + Recall} \tag{13}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{14}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{15}$$

$$IOU = \frac{TP}{(TP + FN + FP)} \tag{16}$$

$$OA = \frac{(TP + TN)}{(TP + TN + FN + FP)} \tag{17}$$

where TP, TN, FP, and FN are true cases, true negative cases, false positive cases, and false negative cases, respectively.

D. Baseline Model

We compare the following 11 baseline models as shown below.

1) FC-EF [14]: This model uses a pixel-level fusion approach that combines images from two different time points into a single input. The combined input is then processed by a fully convolutional neural network.

2) FC-Siam-Conc [14]: This model uses a feature-level fusion approach that incorporates a Siamese U-Net to extract features at multiple levels and fuses the information from the diachronic states by connecting these features in one dimension.

3) FC-Siam-Di [14]: This model uses a feature-level fusion method, employing a Siamese U-Net to extract features at multiple levels and integrating the information of the diachronic state by comparing the differences between these features as a basis.

4) FCN-PP [15]: This model uses a fully convolutional network with pyramidal pooling to effectively analyze the surroundings of a bi-chronological remote sensing image through multiple convolutions. This approach achieves a good balance between expanding the perceptual range and leveraging the context to its fullest extent. The method demonstrates superior performance in terms of automation, noise resistance, improved context utilization, and overcoming the limitations of global pooling of classical methods.

5) DTCDSCN [12]: This model uses a multi-scale feature splicing approach, which produces more distinctive features by incorporating a channel-attention mechanism and a spatial-attention mechanism into the Siamese FCN. In addition, supervised learning is used to ensure the quality of the labeled graph output by the network.

6) IFNet [19]: This model employs a multi-scale feature cascading strategy that applies channel and spatial attention mechanisms at each level of the decoder to deal with diachronic features. Additionally, a deep supervision approach is used to compute the loss at different levels of the decoder, enhancing the training of the intermediate layers of the network. This strategy contributes to enhancing the model performance and effectiveness.

7) STANet [30]: This model improves performance by using a self-attentive mechanism to capture the relationships between different temporal and spatial pixels. It also employs a multi-scale sub-region approach to accommodate different object sizes.

8) FDCNN [31]: This model utilizes migration learning to build a two-channel network that can share weights to generate multi-scale and multi-depth feature disparity maps. It also introduces a loss function guided by the magnitude of variation, which uses a small number of pixel-level samples for training to reduce pseudo-variation.

9) SNUNet [32]: This model employs a network that combines the twin network structure and NestedUnet, and employs a channel attention module to extract features with different semantic information, aiming to effectively retain location information and fully consider shallow features. In addition, a deep supervision module is introduced to enhance the differentiation of intermediate features.

10) DSAMNet [33]: This model employs a metric-based deep supervised attention network. It learns directly from the feature extractor through a change decision module to enhance the learning ability of the feature extractor and generate more useful features. In addition, the intermediate hidden layers are better trained through the deep supervised module.

11) USSFCNet [34]: This model uses an efficient and lightweight approach for fusing spatial and spectral features. It uses cyclic multiscale convolution to capture changing object features at various scales, significantly reducing the number of parameters and redundant computations. In addition, a strategy capable of learning 3D features is introduced for more comprehensive feature extraction.

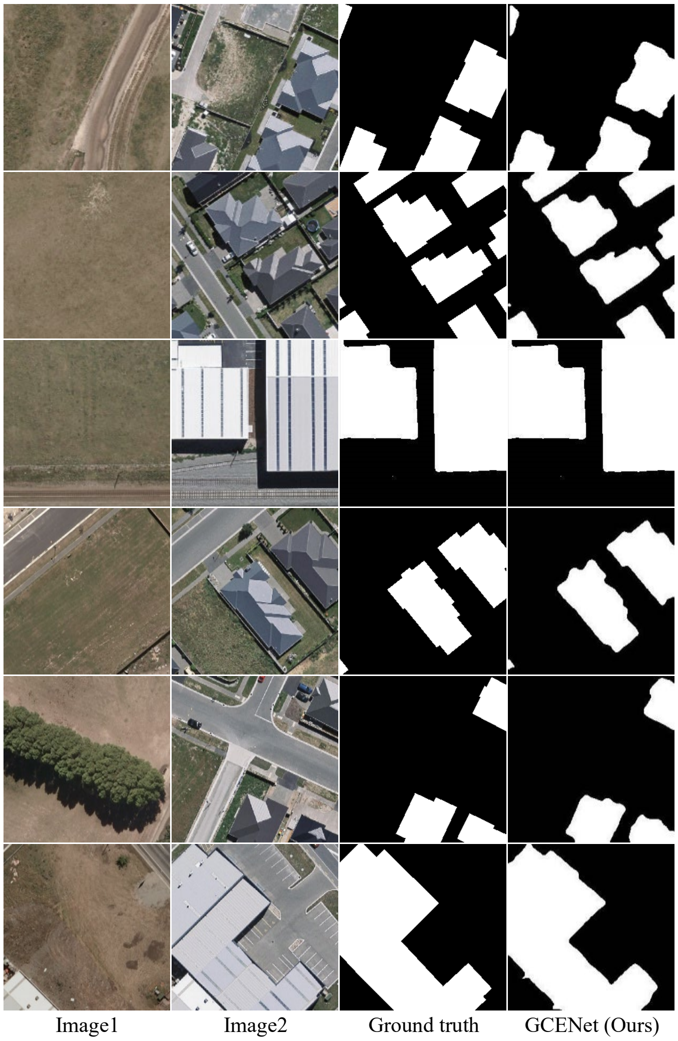


Fig. 7. Visualization of GCENet predictions on the WHU-CD dataset.

12) BIT [24]: This model obtains semantic labels by extracting diachronic image features and then introduces the Transformer encoder-decoder structure, which efficiently models contextual information across the spatio-temporal domain.

E. Comparison Experiment

We tested several baseline models on the WHU-CD and SVCD datasets. Our GCENet model achieved an F1 evaluation metric of 89.95% and SOTA in the three main evaluation metrics of F1, IoU, and OA. As shown in Table I, on the WHU-CD dataset, our F1 metrics improved by 20.58, 23.32, and 31.14 points compared to the classical change detection models FC-EF, FC-Siam-Conc, and FC-Siam-DI, respectively. Compared to some of the top models using convolutional methods in recent years, such as DTCDSCN, STANet, SNUNet, IFNet, and DSAMNet, our F1 metrics improved by 18.00, 7.63, 6.45, 6.55, and 3.24 points, respectively. In addition, our F1 metric also improves by 5.97 points compared to the best recent model BIT using the Transform method.

TABLE I
COMPARATIVE RESULTS ON THE WHU-CD DATASET, THE BOLD FONT REPRESENTS THE OPTIMAL VALUE

Network	Pre(%)	Rec(%)	F1(%)	IoU(%)	OA(%)
FC-EF	71.63	62.75	69.37	53.11	97.61
FC-Siam-Conc	60.88	73.58	66.63	49.95	97.04
FC-Siam-DI	47.33	77.66	58.81	41.66	95.63
STANet	79.37	85.50	82.32	69.95	98.52
DTCDSCN	63.92	82.30	71.95	56.19	97.42
SNUNet	85.60	81.49	83.50	71.67	98.71
IFNet	**96.91**	73.19	83.40	71.52	98.83
BIT	86.64	81.48	83.98	72.89	98.75
DSAMNet	83.90	**90.18**	86.71	76.53	99.05
GCENet(ours)	91.53	88.42	**89.95**	**81.74**	**99.21**

TABLE II
COMPARISON RESULTS ON THE SVCD DATASET, THE BOLD FONT REPRESENTS THE OPTIMAL VALUE

Network	Pre(%)	Rec(%)	F1(%)	IoU(%)	OA(%)
FC-EF	52.67	84.20	64.80	-	-
FC-Siam-Conc	44.07	80.44	56.94	-	-
FC-Siam-DI	61.85	76.69	68.48	-	-
FCN-PP	81.69	90.31	85.78	-	-
FDCNN	83.61	91.70	87.47	-	-
SNUNet	90.92	94.75	92.79	-	-
IFNet	85.33	91.76	88.43	-	-
STANet	88.92	90.83	89.86	81.59	97.58
BIT	92.49	91.22	91.85	84.92	98.09
DSAMNet	91.84	94.14	92.97	86.97	98.32
USSFCNet	93.45	**96.08**	94.74	90.02	-
GCENet(ours)	**95.84**	94.58	**95.16**	**90.78**	**98.87**

TABLE III
COMPARISON OF COMPUTATIONAL EFFICIENCY AND EVALUATION METRICS OF DIFFERENT MODELS OF GCENET ON WHU-CD DATASET, THE BOLD FONT REPRESENTS THE OPTIMAL VALUE

Network	F1(%)	Params.(M)	FLOPs(G)
FC-EF	69.37	0.85	3.34
FC-Siam-Conc	66.63	0.85	3.33
FC-Siam-DI	58.81	1.07	4.08
STANet	82.32	16.93	6.58
DTCDSCN	71.95	41.07	7.21
SNUNet	83.50	12.03	27.44
IFNet	83.40	50.71	41.18
BIT	83.98	3.55	4.35
DSAMNet	86.71	16.95	75.29
GCENet(ours)	**89.95**	3.08	8.81

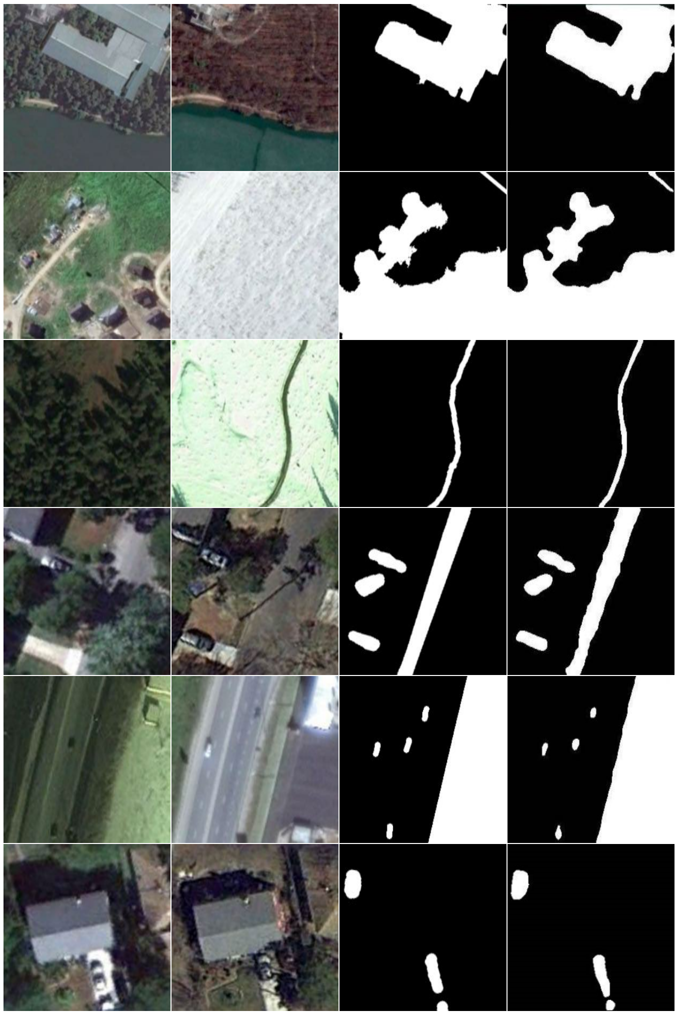


Fig. 8. Visualization of GCENet predictions on the SVCD dataset.

As shown in Table II, on the SVCD dataset, the F1 evaluation metric of our GCENet model achieves 95.16%. In comparison with the classical change detection models FC-EF, FC-Siam-Conc, and FC-Siam-DI, our F1 metrics show improvements of 30.36, 38.22, and 26.68 points, respectively. Furthermore, when compared to some of the recent top models utilizing convolutional methods like FCN-PP, FDCNN, SNUNet, IFNet, STANet, DSAMNet, and USSFCNet, our F1 metrics demonstrate enhancements of 9.38, 38.22, and 26.68 points, respectively. Our F1 metrics improve by 9.38, 7.69, 2.37, 6.73, 5.30, 2.19, and 0.42 points compared to some of the top models in recent years using the convolutional approach, such as FCN-PP, FDCNN, SNUNet, IFNet, STANet, DSAMNet, and USSFCNet, respectively. In addition to our F1 metrics improving by 9.38, 7.69, 2.37, 6.73, 5.30, 2.19, and

0.42 points compared to the best model in recent years using the Transform approach BIT, our F1 metric also improves by 3.31 points.

Fig. 7 and Fig. 8 below show our visualization results on the WHU-CD and SVCD datasets. Our GCENet model performs well on these datasets, and these significant performance improvements are not only shown in terms of F1 metrics, but also visualized on the images. This demonstrates the unique advantages of our model in change detection tasks, providing more accurate results for remote sensing image analysis.

It is worth mentioning that some of these baseline models employ complex network structures such as Feature Pyramid Network (FPN) structures and UNet structures. These structures achieve superior change.

Enhancing detection performance by fusing high-level semantic features with low-level semantic features. Although the backbone network in our GCE-Net model only employs a simple ResNet18 structure, the GCE-Net-based network model is also able to achieve superior performance. This can be attributed to the COTR encoder in our GCE-Net model, which more effectively captures relevant features of the data and enhances the connections between them, thus improving the model's ability to model the relationships between different features.

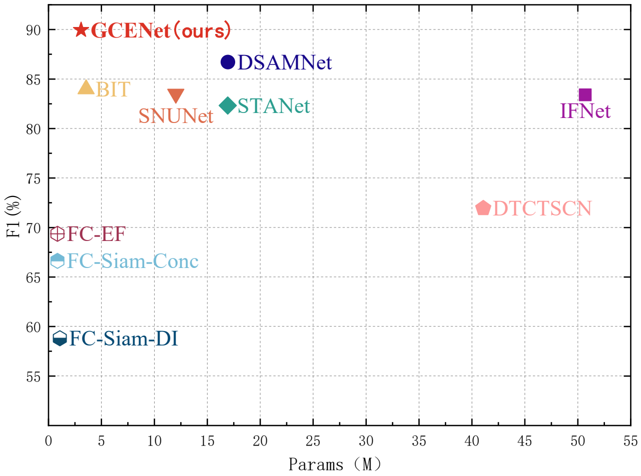


Fig. 9. Comparison of different models in terms of accuracy (F1) and number of model parameters (Params) on the WHU-CD dataset.

We visually tested and analyzed our GCE-Net model, along with other baseline models, for remote sensing image change detection from various perspectives. Our focus was on three key metrics: the number of model parameters, the number of floating-point operations, and the F1 metrics on the WHU-CD dataset. Detailed results are shown in Table III and Fig. 9, comparing our model to the baseline models with fewer model parameters and floating-point operations. Among the baseline models, our method achieves the highest performance in terms of F1 values.

F. Ablation Experiments

In conducting the ablation experiments for the COTR model, we diversified the treatments of Q, K and V in order to deeply investigate their effects on the model performance. Specifically, we introduced the convolution operation for Q, K and V respectively and performed the following different ablation experiments:

1) Convolutional operation on Q only: In this experiment, we perform a convolutional operation on the query information Q only, while keeping the key information K and value information V unchanged. The purpose of this experiment is to evaluate the performance of the model in handling changes in the query information.

2) Convolution operation on K only: In this experiment, we only perform convolution operation to the key information K, while keeping the query information Q and value information V unchanged. This helps understand how the model reacts to changes in key information.

3) Convolution operation on V only: In this experiment, we perform convolution to the value information K only, while keeping query information Q and key information V unchanged. This helps to investigate how the model behaves when dealing with changes in value information.

4) Convolution operations on Q and K while keeping V unchanged: This experiment evaluates the performance of the model when processing both query and key information while keeping the value information unchanged.

5) Convolutional operations on Q and V while keeping K unchanged: This experiment helps to understand how the model performs when processing both query and value information while keeping the key information unchanged.

6) Convolutional operations on K and V while holding Q constant: This experiment helps to investigate how the model behaves when processing both key and value information while keeping the query information constant.

TABLE IV
ABLATION EXPERIMENTS ON THE WHU-CD DATASET WITH OR WITHOUT THE USE OF CONVOLUTION FOR Q, K AND V. THE BOLD FONT REPRESENTS THE OPTIMAL VALUE

Q +conv	K +conv	V +conv	Pre(%)	Rec(%)	F1(%)	IoU(%)	OA(%)
√	×	×	87.41	89.38	88.38	79.19	99.07
×	√	×	88.68	89.87	89.27	80.63	99.14
×	×	√	89.15	89.88	89.51	81.02	99.16
√	√	×	82.54	**90.34**	86.25	75.83	98.85
×	√	√	86.78	89.70	88.21	78.91	99.04
√	×	√	90.45	85.11	87.70	78.09	99.05
√	√	√	**91.53**	88.42	**89.95**	**81.74**	**99.21**

As shown in Table IV, after comparing various ablation experiments, we observe that the model achieves the best performance when using the convolution operation on Q, K and V simultaneously. This result suggests that convolutional operations have a positive combined effect on the COTR model in processing query, key, and value information. These findings provide important insights into the design and performance of the model.

TABLE V

PARAMETER ABLATION OF DECODER LAYERS (E.L) AND ENCODER LAYERS (D.L) ON WHU-CD DATASET, THE BOLD FONT REPRESENTS OPTIMAL VALUES

E.L	D.L	Pre(%)	Rec(%)	F1(%)	IoU(%)	OA(%)
1	1	**91.77**	86.89	89.27	80.62	99.17
1	2	91.53	88.42	**89.95**	**81.74**	**99.21**
1	4	85.07	**90.31**	87.61	77.96	98.99
1	8	84.26	88.49	86.32	75.94	98.88
2	1	86.94	89.29	88.09	78.73	99.04
4	1	85.28	88.44	86.83	76.10	98.97
8	1	86.59	88.67	87.61	78.19	99.08

In examining the effect of the number of layers of encoders and decoders on performance in COTR models. We conducted a series of experiments involving encoders and decoders with different numbers of layers, including 1, 2, 4, and 8 layers. Notably, our study reveals a remarkable result as shown in Table V: the model performance reaches an optimal level when we maintain the COTR encoder at layer 1 and set the Transformer decoder to layer 2. This result has been verified with a high degree of consistency across multiple experiments, highlighting the fact that optimal performance may be achieved with a combination of a shallow encoder and a moderately deep decoder for a given task and dataset. This finding is not only important for optimizing model performance but also helps save computational resources.

V. CONCLUSION

We have proposed GCENet, a global context-enhanced network that aims to address the complexity and performance bottlenecks that exist in the field of change detection in remote sensing images. Through comprehensive experiments and result analysis, GCENet has achieved a significant performance improvement on two change detection datasets. This indicates that GCENet can more accurately recognize changes in remote sensing images, effectively solving the issue where the same object may present different shapes or appearances at different times or spatial locations. Moreover, the relatively low computational complexity of GCENet, while maintaining high performance, makes it an efficient change detection model. When compared to other models, GCENet shows superior efficiency in both the number of parameters and the number of floating-point operations. This implies that GCENet is able to improve the change detection effectiveness in practical applications while maintaining high computational efficiency. In summary, the GCENet proposed in this paper has achieved satisfactory results in the field of remote sensing image change detection, effectively solving the limitations of existing models in complex scenes. Future research directions can include further improving the generalization ability of the model, optimizing computational efficiency, and extending the application to more change detection scenarios. GCENet, as an innovative global context-enhanced network, provides new ideas and solutions for the development of remote sensing image change detection field.

REFERENCES

[1] Ting Bai, Le Wang, Dameng Yin, Kaimin Sun, Yepei Chen, Wenzhuo Li, and Deren Li, “Deep learning for change detection in remote sensing: a review,” Geo-Spatial Information Science, vol. 26, no. 3, pp 262-288, 2023, https://doi.org/10.1080/10095020.2022.2085633.

[2] Wasim Ayub Bagwan, and Ravindra Sopan Gavali, “Dam-triggered Land Use Land Cover change detection and comparison (transition matrix method) of Urmodi River Watershed of Maharashtra, India: a Remote Sensing and GIS approach,” Geology, Ecology, and Landscapes, vol. 7, no. 3, pp 189-197, 2023.

[3] Yongguang Hu, Ali Raza, Neyha Rubab Syed, Siham Acharki, Ram L Ray, Sajjad Hussain, Hossein Dehghanisanij, Muhammad Zubair, and Ahmed Elbeltagi, “Land Use/Land Cover Change Detection and NDVI Estimation in Pakistan’s Southern Punjab Province,” Sustainability, vol. 15, no. 4, pp 3572, 2023.

[4] Fabrice Papa, Jean-François Crétaux, Manuela Grippa, Elodie Robert, Mark Trigg, Raphael M Tshimanga, Benjamin Kitambo, Adrien Paris, Andrew Carr, and Ayan Santos Fleischmann, “Water resources in Africa under global change: monitoring surface waters from space,” Surveys in Geophysics, vol. 44, no. 1, pp 43-93, 2023.

[5] Cheng Gao, Boyao Zhang, Shuaibing Shao, Manqiu Hao, Yuquan Zhang, Yong Xu, Yi Kuang, Lixiang Dong, and Zhuowen Wang, “Risk assessment and zoning of flood disaster in Wuchengxiyu Region, China,” Urban Climate, vol. 49, pp 101562, 2023.

[6] Uwe Stilla, and Yusheng Xu, “Change detection of urban objects using 3D point clouds: A review,” ISPRS Journal of Photogrammetry and Remote Sensing, vol. 197, pp 228-255, 2023.

[7] Yunjia Zou, Ting Shen, Zhengchao Chen, Pan Chen, Xuan Yang, and Luyang Zan, “A Transformer-Based Neural Network with Improved Pyramid Pooling Module for Change Detection in Ecological Redline Monitoring,” Remote Sensing, vol. 15, no. 3, pp 588, 2023.

[8] T. Noulamo, A. Djimeli-Tsajio, J. P. Lienou, and B. Fotsing Talla, “A Multi-Agent Platform for the Remote Monitoring and Diagnostic in Precision Agriculture,” Engineering Letters, vol. 30, no. 3, pp 972-980, 2022.

[9] Song-Bo Zhang, Xiao-Tian Wang, Jie-Sheng Wang, and Xun Liu, "State of Charge Estimation Model for Lithium-ion Batteries Based on Deep Learning Neural Networks," Engineering Letters, vol. 32, no. 2, pp 209-219, 2024.

[10] Shasha Wang, “A Face Recognition Method based on Lightweight Neural Network and Multi Hash Recognition Degree Weighting,” IAENG International Journal of Applied Mathematics, vol. 54, no. 3, pp 581-586, 2024.

[11] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille, “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp 834-848, 2018, https://doi.org/10.1109/TPAMI.2017.2699184.

[12] Yi Liu, Chao Pang, Zongqian Zhan, Xiaomeng Zhang, and Xue 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, 2021, https://doi.org/10.1109/LGRS.2020.2988032.

[13] Huiwei Jiang, Xiangyun Hu, Kun Li, Jinming Zhang, Jinqi Gong, and Mi Zhang, “PGA-SiamNet: Pyramid feature-based attention-guided siamese network for remote sensing orthoimagery building change detection,” Remote Sensing, vol. 12, no. 3, pp 2020, https://doi.org/10.3390/rs12030484.

[14] Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch, "Fully convolutional siamese networks for change detection," Lecture Compilation and indexing terms, Supervised machine learning in Engineering and Computer Science: Proceedings of The World Congress on Engineering Year, 25th IEEE International Conference on Image Processing, ICIP 2018, October 7, 2018 - October 10, 2018, Athens, Greece, pp 4063-4067.

[15] Tao Lei, Yuxiao Zhang, Zhiyong Lv, Shuying Li, Shigang Liu, and Asoke K. Nandi, “Landslide Inventory Mapping From Bitemporal Images Using Deep Convolutional Neural Networks,” IEEE Geoscience and Remote Sensing Letters, vol. 16, no. 6, pp 982-986, 2019, https://doi.org/10.1109/LGRS.2018.2889307.

[16] X. Tang, H. Zhang, L. Mou, F. Liu, X. Zhang, X. X. Zhu, and L. Jiao, “An Unsupervised Remote Sensing Change Detection Method Based on Multiscale Graph Convolutional Network and Metric Learning,” IEEE

Transactions on Geoscience and Remote Sensing, vol. 60, pp 1-15, 2022, https://doi.org/10.1109/TGRS.2021.3106381.

[17] C. Wu, H. Chen, B. Du, and L. Zhang, “Unsupervised Change Detection in Multitemporal VHR Images Based on Deep Kernel PCA Convolutional Mapping Network,” IEEE Transactions on Cybernetics, vol. 52, no. 11, pp 12084-12098, 2022, https://doi.org/10.1109/TCYB.2021.3086884.

[18] F. Shen, Y. Wang, and C. Liu, “Change Detection in SAR Images Based on Improved Non-Subsampled Shearlet Transform and Multi-Scale Feature Fusion CNN,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp 12174-12186, 2021, https://doi.org/10.1109/JSTARS.2021.3126839.

[19] Chenxiao Zhang, Peng Yue, Deodato Tapete, Liangcun Jiang, Boyi Shangguan, Li Huang, and Guangchao 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, https://doi.org/10.1016/j.isprsjprs.2020.06.003.

[20] S. Chen, K. Yang, and R. Stiefelhagen, "DR-TANet: Dynamic Receptive Temporal Attention Network for Street Scene Change Detection," Lecture in Engineering and Computer Science: Proceedings of The World Congress on Engineering Year, 2021 IEEE Intelligent Vehicles Symposium (IV), 11-17 July 2021, Nagoya, Japan, pp 502-509.

[21] T. Lei, J. Wang, H. Ning, X. Wang, D. Xue, Q. Wang, and A. K. Nandi, “Difference Enhancement and Spatial–Spectral Nonlocal Network for Change Detection in VHR Remote Sensing Images,” IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp 1-13, 2022, https://doi.org/10.1109/TGRS.2021.3134691.

[22] Zhuo Zheng, Yanfei Zhong, Shiqi Tian, Ailong Ma, and Liangpei 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.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” Advances in neural information processing systems, vol. 30, pp 2017.

[24] Hao Chen, Zipeng Qi, and Zhenwei Shi, “Remote sensing image change detection with transformers,” IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp 1-14, 2021.

[25] Wele Gedara Chaminda Bandara, and Vishal M. Patel, "A Transformer-Based Siamese Network for Change Detection," Lecture Compilation and indexing terms, Engineering and Computer Science: Proceedings of The World Congress on Engineering Year, 2022 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2022, July 17, 2022 - July 22, 2022, Kuala Lumpur, Malaysia, pp 207-210.

[26] 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," Lecture in Engineering and Computer Science: Proceedings of The World Congress on Engineering Year, 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 10-17 Oct. 2021, Montreal, Canada, pp 9992-10002.

[27] 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, https://doi.org/10.1109/TGRS.2022.3160007.

[28] Y. Dai, T. Zheng, C. Xue, and L. Zhou, “MViT-PCD: A Lightweight ViT-Based Network for Martian Surface Topographic Change Detection,” IEEE Geoscience and Remote Sensing Letters, vol. 20, pp 1-5, 2023, https://doi.org/10.1109/LGRS.2023.3234645.

[29] Y. Feng, H. Xu, J. Jiang, H. Liu, and J. Zheng, “ICIF-Net: Intra-Scale Cross-Interaction and Inter-Scale Feature Fusion Network for Bitemporal Remote Sensing Images Change Detection,” IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp 1-13, 2022, https://doi.org/10.1109/TGRS.2022.3168331.

[30] Hao Chen, and Zhenwei Shi, “A spatial-temporal attention-based method and a new dataset for remote sensing image change detection,” Remote Sensing, vol. 12, no. 10, pp 2020, https://doi.org/10.3390/rs12101662.

[31] Min Zhang, and Wenzhong Shi, “A Feature Difference Convolutional Neural Network-Based Change Detection Method,” IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 10, pp 7232-7246, 2020, https://doi.org/10.1109/TGRS.2020.2981051.

[32] Sheng Fang, Kaiyu Li, Jinyuan Shao, and Zhe Li, “SNUNet-CD: A Densely Connected Siamese Network for Change Detection of VHR Images,” IEEE Geoscience and Remote Sensing Letters, vol. 19, 2022, https://doi.org/10.1109/LGRS.2021.3056416.

[33] Qian Shi, Mengxi Liu, Shengchen Li, Xiaoping Liu, Fei Wang, and Liangpei 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, vol. 60, 2022, https://doi.org/10.1109/TGRS.2021.3085870.

[34] Tao Lei, Xinzhe Geng, Hailong Ning, Zhiyong Lv, Maoguo Gong, Yaochu Jin, and Asoke K Nandi, “Ultralightweight Spatial–Spectral Feature Cooperation Network for Change Detection in Remote Sensing Images,” IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp 1-14, 2023.