Position-aware graph-CNN fusion network: an integrated approach combining geospatial information and graph attention network for multi-class change detection

Moyang Wang, Xiang Li, Kun Tan, *Senior Member, IEEE*, Joseph Mango, Chen Pan and Di Zhang

***Abstract*—Urban change detection is crucial for informed decision-making but faces various challenges, including complex features, rapid changes, and extensive human interventions. These challenges underscore the urgent need for innovative multi-class change detection (MCD) techniques that extensively incorporate deep learning. Despite several successes achieved with the deep learning based MCD methods, still certain shortcomings persist, including the disregard for spatial principles, which significantly hinders the seamless integration of geoscience-knowledge and artificial-intelligence. In this paper, a novel deep learning model known as the Position-aware Graph-CNN Fusion Network (PGCFN) is introduced, integrating spatial position encoding to effectively detect urban changes. The model's first part encodes geospatial positions following Tobler's first law of geography. It then integrates encoded positions into a multi-class change detection model, combining a graph attention network with a convolutional neural network to enhance performance. The model was tested on 0.5-meter resolution remote sensing images, achieving an impressive minimum Mean Intersection over Union (MIoU) score of 91.20%. Additionally, the model's position-aware graph attention module exhibited a strong emphasis on geographic-proximity when evaluating connections between superpixels. Overall, these findings affirm that our model could effectively addresses urban change detection challenges and significantly enhances the integration of geoscience knowledge and artificial intelligence.**

***Index Terms*—Geospatial artificial intelligence, multi-class change detection, graph attention network, position information encoding, urban changes.**

I. INTRODUCTION

Remote sensing image change detection plays a pivotal role in the field of geospatial artificial intelligence (GeoAI)[1]. It involves the technology of identifying and analyzing alterations occurring among geographical features within a same geographic area between different time periods [2-4]. Over recent decades, the field of change detection (CD) has evolved and developed into a widely researched and applied area, influencing domains such as natural resource management, urban planning and disaster monitoring [5-7].

The continuous advancement of artificial satellite technology and sensor capabilities has enabled the rapid acquisition of extensive remote sensing imagery data, providing robust support for dynamic land cover change analysis [8]. Spatial resolutions in optical remote sensing images have significantly improved from hundreds or tens of meters in the past (such as MODIS and Landsat series satellites) to the current meter and sub-meter levels (such as GaoFen-2 satellite) [9, 10]. Although high-resolution images contain rich spatial structures and more refined texture and morphological information, they also suffer from increased noise due to excessive spatial resolution [11]. Additionally, a substantial amount of original data remains incorrectly labeled, presenting a significant challenge for subsequent interpretation and application of remote sensing images [12, 13]. Early binary change detection (BCD) methods were focused on to detecting changes and unchanged areas between dual-temporal images but faced practical implementation problems. In recent years, multi-class change detection (MCD), capable of distinguishing different change categories, has emerged as a prominent area of academic research [14]. One straightforward approach to MCD involves classifying images from different times individually, and then comparing those classifications to generate multi-class change maps, commonly known as post classification method (PCC) [15]. However, this method often suffers from low detection accuracy due to error accumulation. Another strategy is the direct classification method (DC), which is the opposite of the PCC method [2]. DC regards MCD as a multi temporal image classification task, directly classifying the processed multi temporal images, where each change type is considered as a

Manuscript received xx 2023. This work was supported in part by the International Research Center of Big Data for Sustainable Development Goals under Grant CBAS2022GSP07, in part by the Chongqing Natural Science Foundation under Grant CSTB2022NSCQMSX2069, and in part by the Ministry of Education of China under Grant 19JZD023. (*Corresponding author: Xiang Li*).

Xiang Li is with the Key Laboratory of Geographic Information Science (Ministry of Education), the Shanghai Key Lab for Urban Ecological Processes and Eco-Restoration, and the Key Laboratory of Spatial-Temporal Big Data Analysis and Application of Natural Resources in Megacities (Ministry of Natural Resources), East China Normal University, Shanghai 200241, China (e-mail: xli@geo.ecnu.edu.cn).

Moyang Wang, Kun Tan and Di Zhang are with the Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai 200241, China (e-mail: wangx_teresa@stu.ecnu.edu.cn; tankuncu@gmail.com; zhangdee@stu.ecnu.edu.cn).

Joseph Mango is with the Department of Transportation and Geotechnical Engineering, University of Dar es Salaam, P.O.Box 35131, Dar es Salaam, Tanzania (e-mail: jsmmango@udsm.ac.tz).

Chen Pan is with the Shanghai Municipal Institute of Surveying and Mapping, Shanghai 200063, China (e-mail: panpan_tj@126.com).

separate class. Compared with PCC, DC achieves higher detection accuracy as it doesn't accumulate errors. However, it requires a substantial number of manually labeled samples of various changes, which increases the complexity of classification to some extent.

Currently, deep learning (DL)-based methods have seen efficient use in various fields, such as image semantic segmentation [16], object detection [17, 18], and change detection [19, 20]. Similar to traditional methods, DL-based MCD methods can also be categorized into PCC and DC [21]. These methods typically employ semantic segmentation techniques combined with various neural networks for classification, with convolutional neural networks (CNNs) and their variants being the most prevalent choices [22-24]. MCD models typically use dual temporal images as inputs, with each pixel assigned a unique semantic change label. The latest representative deep learning network frameworks used for MCD are dual branch network frameworks like ReCNN [25], Bi-SRNet [26], and Y-Net [27]. In addition to CNNs, graph neural networks (GNNs) with their unique graph-based structure, are suitable for analyzing geographical data, as they can handle irregular shapes. One limitation of existing GNN-based models is their uniform treatment of relations and reliance on fixed network parameters, often overlooking valuable information carried by individual relations. The literature revels that this challenge could be solved with another graph-based network called the graph attention network (GAT). With GAT, each node in the graph can be assigned varying weights based on the attributes of its neighboring nodes, and these weights are updated iteratively. This attention mechanism enhances the model's ability to capture spatial information relevance. Consequently, various variants of GNNs has garnered increasing attention in the domain of MCD [20].

Despite the promising future, there are still several common issues in the design of deep learning models, including the neglect of domain knowledge, practical experience, and established scientific principles such as Tobler's First Law of Geography (TFL) [28-30]. In geography, the concept of spatial regularity demonstrated with the Tobler's First Law (TFL) remains steadfast in any research pertaining to spatial and locational analysis: "*Everything is related to everything else, but near things are more related than distant things.*". TFL has provided valuable guidance in the development of specific methods, such as the Geographically Weighted Regression model (GWR). Additionally, concepts like spatial autocorrelation and edge features have been integrated as prior geographic spatial knowledge within deep learning models, particularly in the context of weak supervision terrain detection [30]. Li et al. [31] introduced the notion of classification-based reasoning, which involves improving the classification outputs of deep learning modules through the utilization of ontological reasoning rules. This approach enhances the model's ability to distinguish objects with spatial similarities in remote sensing images. Taken together, the theories and methods developed to-date expand current deep learning techniques to incorporate spatially explicit models and, in that way, they are enhancing the adaptability of artificial intelligence in geospatial domains while also improving interpretability. However, their explicit utilization in multi-class change detection remains unexplored.

In this study, we focus on the powerful synergy between deep learning and geospatial knowledge to advance multi-class change detection model by incorporating fundamental spatial theories. Specifically, we introduce a novel supervised learning MCD framework called the Position-aware Graph-CNN Fusion Network (PGCFN), leveraging a graph attention network within the context of spatial distance perception. This study develops a position information encoding mechanism that employs the graph attention network to simulate diverse spatial distance relationships among objects. Furthermore, we enhance the graph by incorporating relative spatial distance information into its edges through spatial information encoding. To improve the overall performance of the detection model, we use a pixel-superpixel mapping matrix to facilitate seamless feature propagation between image pixels and graph nodes. This enables our model to learn feature information at various scales. Lastly, we implement our proposed model to appreciate its applicability and accuracy using high-resolution satellite imagery. The key contributions associated with the development of the PGCFN are as follows:

1) We propose a novel position information encoding mechanism by combing the order matrix and the two-dimensional spatial distance between superpixels. The order matrix consists of the k-order adjacent relations between graph nodes;
2) We present a new graph attention model incorporating spatial position information (P-GAT) to autonomously learn spatial relations of objects under the guidance of spatial knowledge;
3) The incorporation of position encoding into the model enhances its ability to attain a more intricate and specific understanding of the geospatial structure. This refinement proves beneficial in optimizing the overall performance of the model.

The rest of this paper is organized as follows: first, we present an overview of the current state of research in MCD. Then, we delve into the integration of spatial position information into our model design, followed by a detailed explanation of our approach. Finally, we outline the experimental setup and results, and engage in a discussion of potential directions for future research.

II. Related Work

As our research primarily focuses on amalgamating spatial principle with deep learning models via the development of novel GNN structures, we will provide a concise overview of deep learning-based MCD and the integration of geospatial knowledge and deep learning models.

A. Deep learning-based change detection

With the development of deep learning theory, methods based on deep learning are gradually being applied to BCD and MCD tasks. Early deep learning methods were mainly based on the PCC method that often uses different semantic segmentation deep learning networks for classification. After classification, multiple variations were obtained through comparison. The deep research of direct classification methods obtains not only changes and non-changes of urban land cover through threshold segmentation, but also subdivides their observed phenomena into various categories through clustering and classification [21]. Convolutional neural networks are widely used in CD tasks, owing to their capacity to directly segment multiple temporal images and to model dynamic information by combining and transforming various temporal features through stacked convolutional layers. Daudt et al. [2] proposed a multi-task framework that includes a fully convolutional network for BCD, in addition to a fully convolutional network branch for classification. Ding et al. [26] designed a dual-temporal semantic reasoning network to infer the semantic correlation between single-temporal and cross-temporal phases, and employed a novel loss function to enhance the semantic consistency of change detection outcomes. Recently, the foundation model has garnered significant attention. In [32], a remote sensing foundation model for spectral data was introduced. This model incorporated convolution operations and possessed the capability to learn intrinsic knowledge representations, thereby offering valuable insights for comprehending various downstream applications in remote sensing, notably in the domain of change detection.

These CD models typically take dual temporal images as input and subsequently generate pixel wise "from-to" change maps, wherein each pixel is associated with a distinct encoded semantic change label [19, 33]. Although existing methods based on CNNs have been proven to provide effective results in many cases, there are still some limitations. Specifically, ground objects in remote sensing images often have multiple scales and shapes, while the convolution kernel in CNNs primarily conduct convolutions within regular rectangular regions, making it difficult to comprehensively capture internal correlations between adjacent objects [34]. Recently, graph convolutional networks (GCNs), as an extension of GNNs, have received an increasing attention owing to their ability to perform convolution operations on graphs with arbitrary structures. GCNs exhibit a unique graph-based structure, enabling them to break the constraints of regular shapes, making them suitable for application in geographic data analysis.

Saha et al. [35], for instance, utilized GNN to improve change detection performance and proposed a new graph construction method to process segmented objects into graph representations that can be processed by GCN to optimize their loss functions on labeled objects only. The iterative training method helps to propagate label information from labeled nodes to unlabeled nodes, thereby detecting changes in unlabeled data. Another study by Zhou et al. [20] applied graph convolution to the MCD task and designed a Siamese graph convolutional network (SIGNet) inspired by twin-structure. It employed the cross-attention mechanism to establish semantic connections with the category information within the dataset during spatial relationship reasoning. The recent study by Liu et al. [36], has demonstrated that the integration of CNN and GCN can effectively capture spatial topological relationships, thereby enhancing the robustness of image classification models.

With the increasing complexity of algorithms, researchers are increasingly pushing for the development of larger deep learning models, which stems from the desire to improve accuracy and performance for more complex prediction tasks. However, the neglect of inherent spatial laws often occurs in the process of DL model design, hindering further research on the interpretability of DL models. The focus of many DL models in the MCD tasks is to design network structure to enhance the transmission of information between semantic features and corresponding labels. In contrast, our work considers the guiding role of geographic spatial principle in model design, aiming to improve detection accuracy while simultaneously investigating the potential impact of spatial principle on deep learning models.

B. Geospatial analysis method with deep learning

Geography provides a unified perspective for comprehending the world and society, guided by well-established theories such as the Tobler's first law (TFL) [37]. It suggests that objects or phenomena that are geographically close to each other are more likely to be similar or have spatial relationship compared to objects that are farther apart. TFL has played a pivotal role in shaping the design of certain spatial techniques, such as geographic weighted regression (GWR), which is a localized linear regression method rooted in modeling spatial change relationships. Similar ideas these days have been employed with deep learning to produce different robust results. As a subset of artificial intelligence (AI), deep learning represents a significant advancement of models from shallow to deep architectures, allowing complex patterns to be modeled and extracted by utilizing artificial neural networks.

The integration of AI and geospatial reasoning can amplify the interpretability of models and facilitate a more contextually suitable adaptation of AI to the geospatial domain [1, 38]. Knowledge and spatial principles from the geospatial domain have been explored to provide guidance in crafting more effective deep learning models. Research, such as of the Julian et al. [39], combines artificial neural networks and geographical weighting method to model spatial heterogeneous relationships. By incorporating geographically weighted learning into the neural network, models learn parameters that vary based on location, rather than assuming a uniform value. Based on the principle of TFL, Li et al. [30] transformed two-dimensional images into one-dimensional sequence data, preserving the inherent spatial continuity of the original data. Subsequently, they developed an enhanced LSTM model capable for processing 1D sequences and object localization. Unlike the general object detection models, this method does not need bounding box labels.

Images, especially remote sensing or geospatial images, often exhibit spatial autocorrelation, meaning that nearby pixels or regions in the image are more likely to have similar values or characteristics than those farther apart. In computer vision tasks like object detection and image segmentation, understanding spatial relationships between objects or regions within an image is crucial. With the integration of geospatial knowledge and spatial principles, models can comprehend intrinsic spatial relationships among ground objects to construct a more efficient deep learning model, thereby advancing the field of interpretability research. Nevertheless, at present, there are few related studies about geospatial-aware deep learning in the domain of MCD on remote sensing images.

III. Materials and Method

Essentially, Tobler's First Law serves as a perfect description of spatial autocorrelation, which highlights internal relationships between geographical entities. There is a close relationship between spatial distance and internal relationships between entities. How to deeply integrate spatial theory and laws is a key aspect of our proposed model design. In this section, we introduce a position-aware graph attention module to address the multi-class change detection task, designed in alignment with Tobler's First Law.

A. Position information encoding

Unlike general graphs, such as social network data and recommendation systems, geospatial graph data has a special spatial structure (two-dimensional or three-dimensional). This may have an impact on the interactions between geographical entities. The ability to learn this spatial structure and position information is crucial for exploring the interpretability of models. To better integrate geospatial information and topological structure of ground objects, a novel position-aware graph attention module was designed.

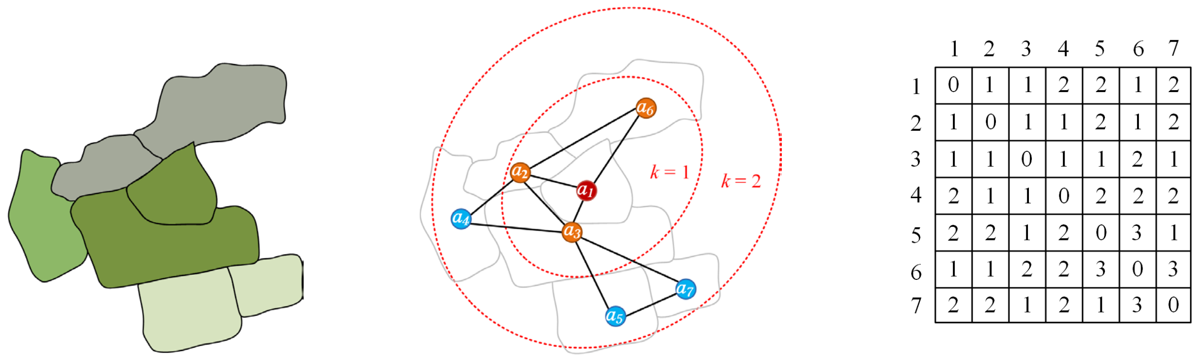
First it is important to clarify the construction method of the graph. Although the nodes in the graph can be represented by image pixels, this will result in a huge graph and the calculation will be difficult. To take full advantage of the learning power of GNNs, a common approach is to segment the image into larger parts, which are often referred to as superpixels. Each superpixel is treated as a node within the graph, and the connections between these superpixels depict their interactions. This segmentation process serves to diminish the graph's scale, rendering calculations and learning procedures more efficient. To begin with, we employed SLIC [40] to partition the image into superpixels, representing the objects in the image. In our study, two temporal images were segmented separately and merged to obtain a set of superpixels, i.e., $S = \{S_i\}_{i=1}^{N}$, where S_i represents the i-th superpixel, with N being the number of superpixels. Like Liu et al. [36], we represented each superpixel with its centroid as a node. Finally, the image was transformed into an undirected graph $G = (\nu, \varepsilon)$ by creating adjacency relationships among superpixels, where ν and ε represent nodes and edges, respectively.

In the field of natural language processing, position encoding plays a vital role by assigning vectors that represent the positional information of elements within sequential data. Spatial distances carry distinct meanings in correlation with ground objects, and positional encoding can be employed to represent the spatial distance between objects in geographic space. This encoding method offers contextual information within neural networks, enabling the network to discern the significance of inputs originating from diverse locations.

In our study, the positions of the ground object were defined in terms of 2D coordinates firstly, forming an initial position matrix $\boldsymbol{I} \in R^{N\times 2}$, where N represents the number of ground objects. Considering the variability of artificially defined coordinates, we transformed $\boldsymbol{I}$ into a relative spatial matrix, namely the distance matrix $\boldsymbol{D} \in R^{N\times N}$. Among them, the adjacency of ground objects is one of the important properties in graphs, and subgraphs of k-th order allow to observe the relevance of ground objects from different levels. As depicted in the Fig. 1(c), for any given node n, there exists a set consisting of N-1 nodes, where the nodes in this set share a k-th order adjacency relationship with node n. Based on this premise, we constructed the order matrix $\boldsymbol{K} \in Z^{N\times N}$. From (1), the order was converted into a part of the distance parameter, resulting in a position relationship matrix $\boldsymbol{D'} \in R^{N\times N}$.

$$\boldsymbol{D}'_{i,j} = 1/(\frac{1}{\boldsymbol{K}_{i,j}+1} * \boldsymbol{D}_{i,j})^{r} \tag{1}$$

where r is the exponent determining the rate of weight decay with distance. To represent spatial information, the scalar distance was



(a)

(b)

	1	2	3	4	5	6	7
1	0	1	1	2	2	1	2
2	1	0	1	1	2	1	2
3	1	1	0	1	1	2	1
4	2	1	1	0	2	2	2
5	2	2	1	2	0	3	1
6	1	1	2	2	3	0	3
7	2	2	1	2	1	3	0

(c)

Fig. 1. Visual illustration of the k-th order neighborhood in graph. (a) An example of ground objects. (b) Simple neighborhood. (c) Order matrix.

divided into l components, and this information was encoded by utilizing a one-hot encoder, resulting in a multi-dimensional relational matrix $\boldsymbol{P} \in Z^{N \times N \times l}$. Illustrated in Fig. 2, we partitioned the neighbors of a_1 into three different spatial relations, i.e. l = 3. Subsequently, a learnable linear transformation was employed to obtain the position embedding $\boldsymbol{D}_{i,j}^{R}$ for each node.

$$\boldsymbol{D}_{i.j}^{R} = \boldsymbol{W}_p \boldsymbol{P}_{i,j} \tag{2}$$

where $\boldsymbol{W}_p \in R^{d \times l}$ represents a weight matrix learned by the network. Next, we integrated the position embedding into the graph attention layer at the node-level.

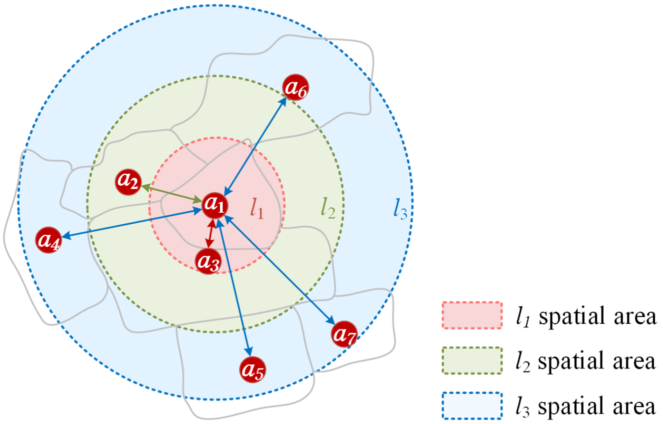


Fig. 2. Spatial position encoding.

B. Position-aware graph attention network

The attention mechanism was originally developed for natural language processing and has been used in a wide range of different applications. In RS image processing based on deep learning, attention mechanism is typically employed on particular feature layers or a specific region of an image [41]. Previous studies have provided evidence of the favorable influence of the attention mechanism on the performance of deep learning methods in RS image processing [42]. According to TFL, it points out that nearby objects exhibit stronger correlations with each other, and the attention mechanism can be guided to facilitate the model in autonomously learning the spatial relationships among these objects. The core concept of the graph attention network is to assign an attention weight to each node, guiding the propagation and aggregation of information [43]. Here, we introduce the building block layers for constructing the distance-aware graph attention network module, starting with the description of a single graph attention layer as follows.

Given N node features $\boldsymbol{h} = \left\{ \vec{h}_1, \vec{h}_2, \cdots, \vec{h}_N \right\}, \vec{h}_i \in R^F$, where F represents the feature numbers in each node. As an initial step, the node features are first dimensionally extended through a shared parameter matrix, $\boldsymbol{W} \in R^{F' \times F}$. We propose the position-aware attention to learn the weight among multi-relation objects. The importance of node j towards node i can be formulated as,

$$e_{i,j} = atten\left(h_i, h_j\right) = \boldsymbol{a}^T \left[\boldsymbol{W}\vec{h}_i \middle\| \boldsymbol{W}\vec{h}_j \right] \boldsymbol{D}_{i.j}^{R} \tag{3}$$

where ‖ represents the concatenation operation, and a shared attention mechanism $\boldsymbol{a} \in R^{F' \times F'}$ is applied to map the concatenated high-dimensional features into real numbers. Then the softmax function is used for normalization:

$$\alpha_{i.j} = softmax(e_{i,j}) = \frac{exp(e_{i,j})}{\sum_{k \in \mathbf{N}_i} e_{i,k}} \tag{4}$$

where $\mathbf{N}_i$ is the first-order neighbors of node i in the graph. The LeakyReLU nonlinearity is applied as the activation function. When expanded completely, the coefficients computed by the attention mechanism can be expressed as:

$$\alpha_{i,j} \in \frac{exp(\text{LeakyReLU}(\boldsymbol{a}^T \left[\boldsymbol{W}\vec{h}_i \middle\| \boldsymbol{W}\vec{h}_j \right] \boldsymbol{D}_{i.j}^{R}))}{\sum_{k \in \mathbf{N}_i} exp(\text{LeakyReLU}(\boldsymbol{a}^T \left[\boldsymbol{W}\vec{h}_i \middle\| \boldsymbol{W}\vec{h}_k \right] \boldsymbol{D}_{i.k}^{R}))} \tag{5}$$

Once obtained, the normalized attention coefficient is utilized to calculate the linear combination of corresponding features. After that, a nonlinearity σ is applied to obtain the final output feature of each node.

$$\vec{h}_i' = \sigma(\sum_{j \in \mathbf{N}_i} \alpha_{i,j} \boldsymbol{W}\vec{h}_j) \tag{6}$$

To enhance stability of the attention-based learning process, a multi-head attention is similarly employed as by [43] to extending our mechanism. Here, K independent attention mechanisms are employed to execute the transformation outlined in (6), and then connect their features to produce the following output feature representation:

$$\vec{h}_i' = \Big\|_{k=1}^{K} \sigma(\sum_{j \in \mathbf{N}_i} \alpha_{i,j}^{k} \boldsymbol{W}^k \vec{h}_j) \tag{7}$$

where ‖ denotes concatenation, $\alpha_{i,j}^{k}$ represents the normalized attention coefficients calculated by the k-th position-aware attention mechanism, and $\boldsymbol{W}^k$ is the weight matrix associated with the corresponding input.

By using a specially designed position-aware attention mechanism, all relevant neighbors of each node were aggregated, so that the combination of position representations contained necessary spatial and topological information. Notably, position information encoding in this study intentionally injected geospatial information into the deep learning model. The primary goal is to enhance the model's understanding of geographic location and spatial relationships, making it more suitable for geospatial analysis tasks. Based on the design of the distance-aware attention layer, a change detection model was designed.

C. Multi-class change detection model architecture

Based on the position-aware graph attention module (P-GAT), a deep learning network for multi-class change detection is proposed with high-resolution images, as shown in Fig. 3. Integrating CNNs and GNNs will serve to enhance and enrich each other's respective strengths, therefore, our detection model integrated CNN and P-GAT for pixel-level and object-level feature fusion using an association matrix.

We represent the concatenated bi-temporal remote sensing

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3350573

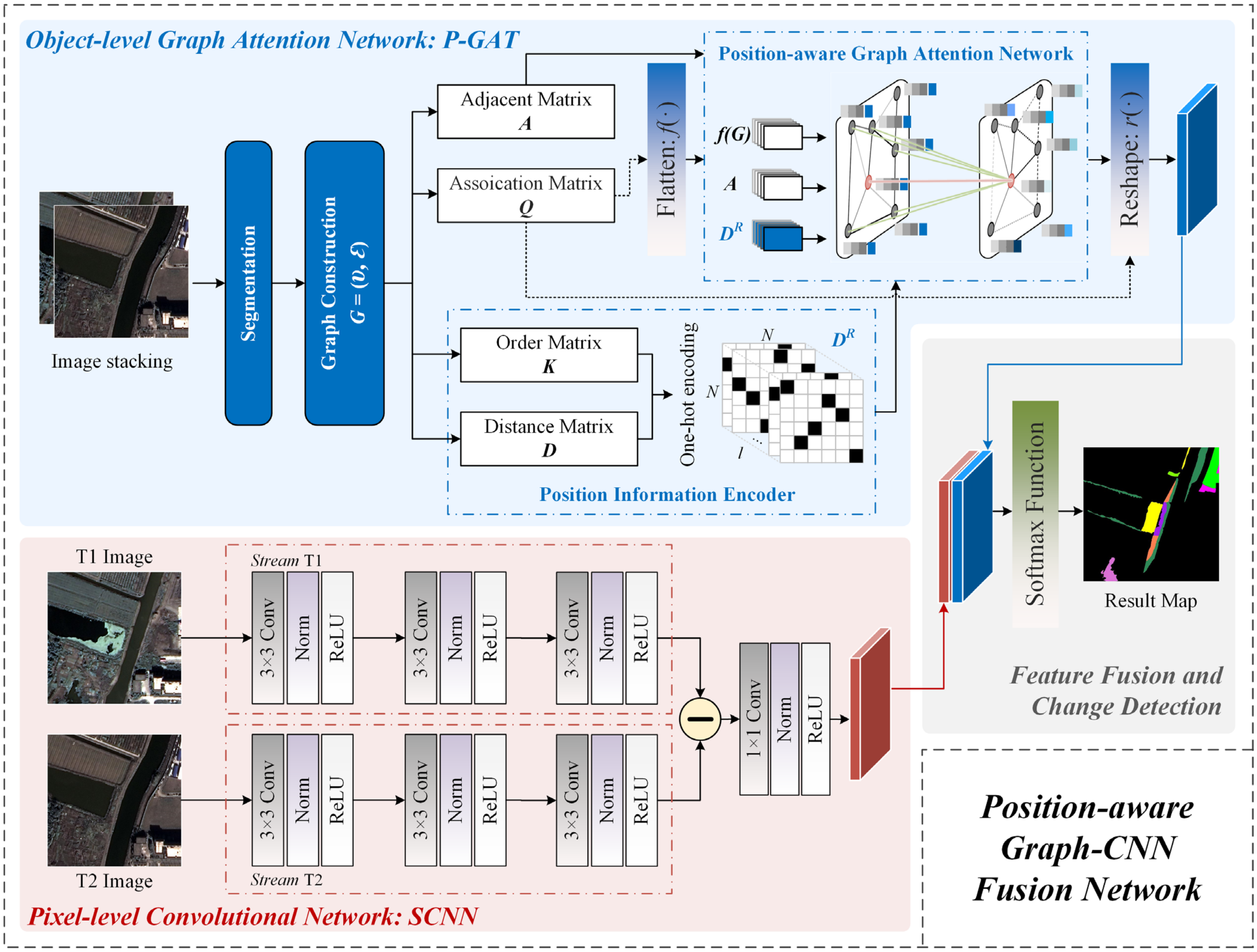


Fig. 3. Illustration of the proposed PGCFN model.

(RS) images as $X=(X_1^{(h,w,c)}, X_2^{(h,w,c)})$, where $X_1^{(h,w,c)}$ and $X_2^{(h,w,c)}$ correspond to the images captured at two different time points, each having dimensions (*h*, *w*, *c*). In this notation, *h* represents height, *w* represents width, and *c* represents the channel size.

As shown in Fig. 3, the pixel-level convolutional network (SCNN) designed is composed of two parallel multi-layer convolutional modules. Each stream consists of three groups of conventional convolutions. In each group, there are two convolutional layers to transform the original space and spectral features into the high-dimensional feature space. After progressive abstraction through stacked convolutional layers, the deepest layer in streams T1 and T2 captures compact local information. The pixel-level difference features are then obtained through a difference operation. Let H^{pixel} represents the output feature map after SCNN and we can express it as:

$$H^{pixel} = Scnn(X_1, X_2) \tag{8}$$

Utilizing position-aware attention, we constructed a spatial position guided attention module based on superpixels. This module is integrated into a two-layer P-GAT model. The first layer comprises four attention heads, while the second layer is dedicated to feature combination and classification, employing a single attention head, followed by an exponential linear unit (ELU) nonlinearity. The output feature after P-GAT, denoted as H^{object} is:

$$H^{object} = Pgat(X) \tag{9}$$

After the segmentation, *N* superpixels are obtained. $S=\{S_i\}_{i=1}^{N}$ is the superpixel set, $S_i=\{x_j^i\}_{j=1}^{N}$ denotes the *i*-th superpixel, and x_j^i represents the *j*-th pixel in *i*-th superpixel. Firstly, let Q be the association matrix between pixels and superpixels, as shown in (10).

$$Q_{i,j}=\begin{cases}0, & otherwise\\ 1, & if\ \hat{X}_i \in S_j\end{cases} \tag{10}$$

$$\hat{X} = flatten(X) \tag{11}$$

where $\hat{X}_i$ is the *i*-th pixel in X, *flatten* (·) is flattening data according to spatial dimensions. Next, we can quickly encode and decode the image into graph nodes through matrix operations by (12) and (13), respectively.

$$V = Q^T flatten(X) = Q^T \hat{X} \tag{12}$$

$$X' = reshape(QV) \tag{13}$$

where $\boldsymbol{V}$ is the feature vector of the graph nodes, and *reshape* $(\cdot)$ function is applied to restore the spatial dimension of the flattened data. After completing the decoding process, the graphic features can be projected back into the image space[36].

Our method defines two subnetworks, SCNN and P-GAT. Specifically, SCNN is responsible for extracting dual-branch difference features and local spectral spatial features at the pixel level. On the other hand, P-GAT is employed to learn data representation related to objects and position features, thereby generating superpixel features. Ultimately, these two different levels of integrated and utilized for change detection. The entire feature fusion process can be expressed as:

$$\boldsymbol{F}_{map} = reshape(\boldsymbol{Q}(Pgat(\boldsymbol{Q}^T flatten(\boldsymbol{X})))) \| Scnn(\boldsymbol{X}_1, \boldsymbol{X}_2) \quad (14)$$

Finally, to generate the probability for each change class on each pixel, a softmax classifier is employed to classify the feature map $\boldsymbol{F}_{map}$. As a result, the final change detection map $\boldsymbol{Y}$ is defined as

$$\boldsymbol{Y} = softmax(Linear(\boldsymbol{F}_{map})) \quad (15)$$

where $Linear(\cdot)$ denotes a fully connected layer, and $softmax(\cdot)$ is a softmax classifier.

IV. Experimental Results

A. Dataset description

This study conducted an experiment to appreciate our designed model using high-resolution satellite images taken in Pudong New Area, Shanghai, China (Fig. 4). The acquisition times of the two SuperView-1 (SV-1) images were in October 2018 and November 2021, respectively, with a spatial resolution of 0.5 meters. Both sets of data have an image size of 1000*1000 pixels, consisting of four bands: blue, green, red, and near-infrared. For land cover classification tailored to the specific conditions of the experimental area, we designed a system with five categories: built-up areas (including residential, commercial, industrial and service zones, roads, and other mixed development zones), water (including rivers, streams, ponds, and lakes), farmland (including agricultural fields, planting greenhouse, orchards, and fallow lands), green land (including forests, shrubs, lawns, mixed wooded areas, etc.), and bare land (including undeveloped lands, exposed soils and unused lands).

The first dataset comprises 11 change categories, of which the top three categories are the change of Farmland to Built-up areas, Water, and Green land to Built-up areas. Notably, the category with the largest proportion has 75,318 pixels, while the category with the smallest proportion of Built-up areas to Water contains only 920 pixels, highlighting a significant class imbalance. The second dataset contains four types of changes, namely, Green land to Built-up areas and Water, and Built-up areas to Water and Green land, respectively. Remarkably, the pixel counts for all categories are evenly distributed. The last dataset comprises five distinct change categories: change from Bare land to Water and Green land, from Water to Green land and Bare land, and from Farmland to Bare land. Among them, the change from Farmland to Bare land is only 1,646 pixels, accounting for the smallest proportion. There are 59,817 pixels from Water to Green land, accounting for the largest proportion. Importantly, these three datasets do not share identical change information, and the proportion of each type varies significantly, highlighting the challenges associated with MCD. Fig. 5 shows the true-color images and reference change maps for the three datasets.

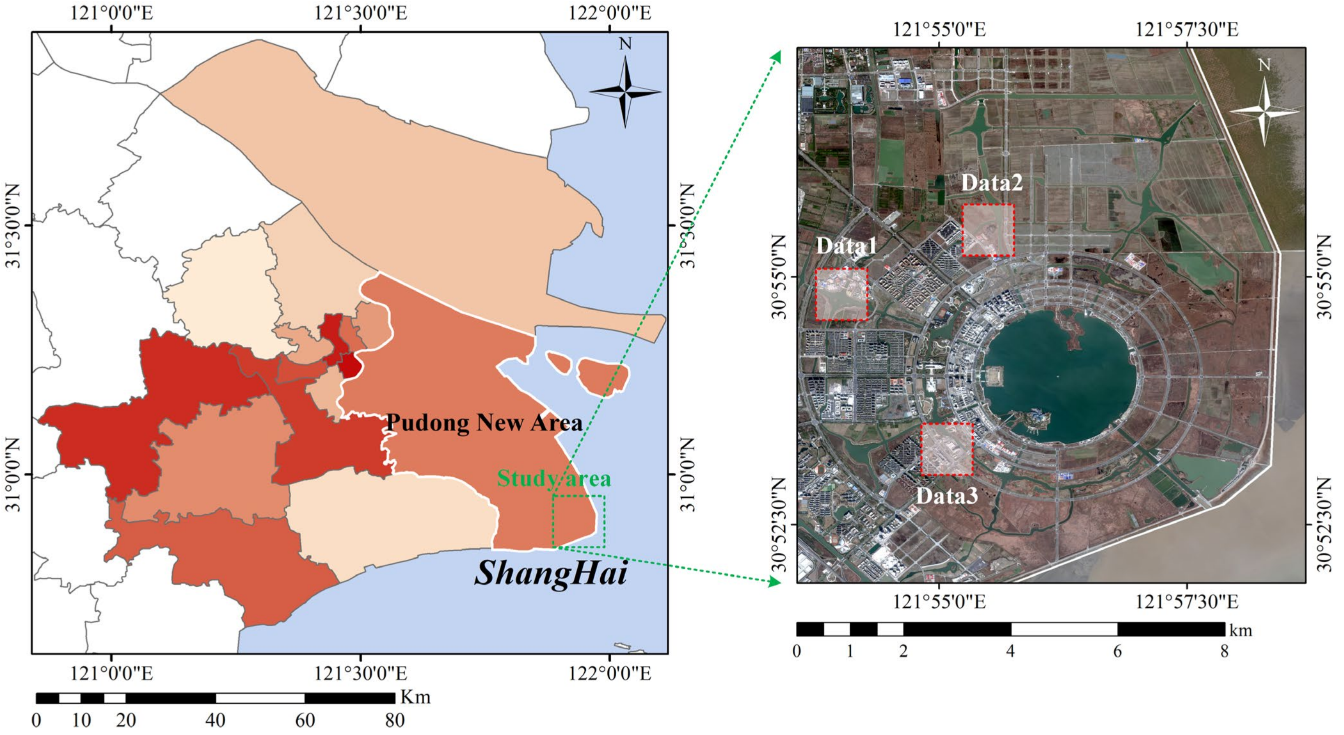


Fig. 4. Illustration of the study area.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3350573

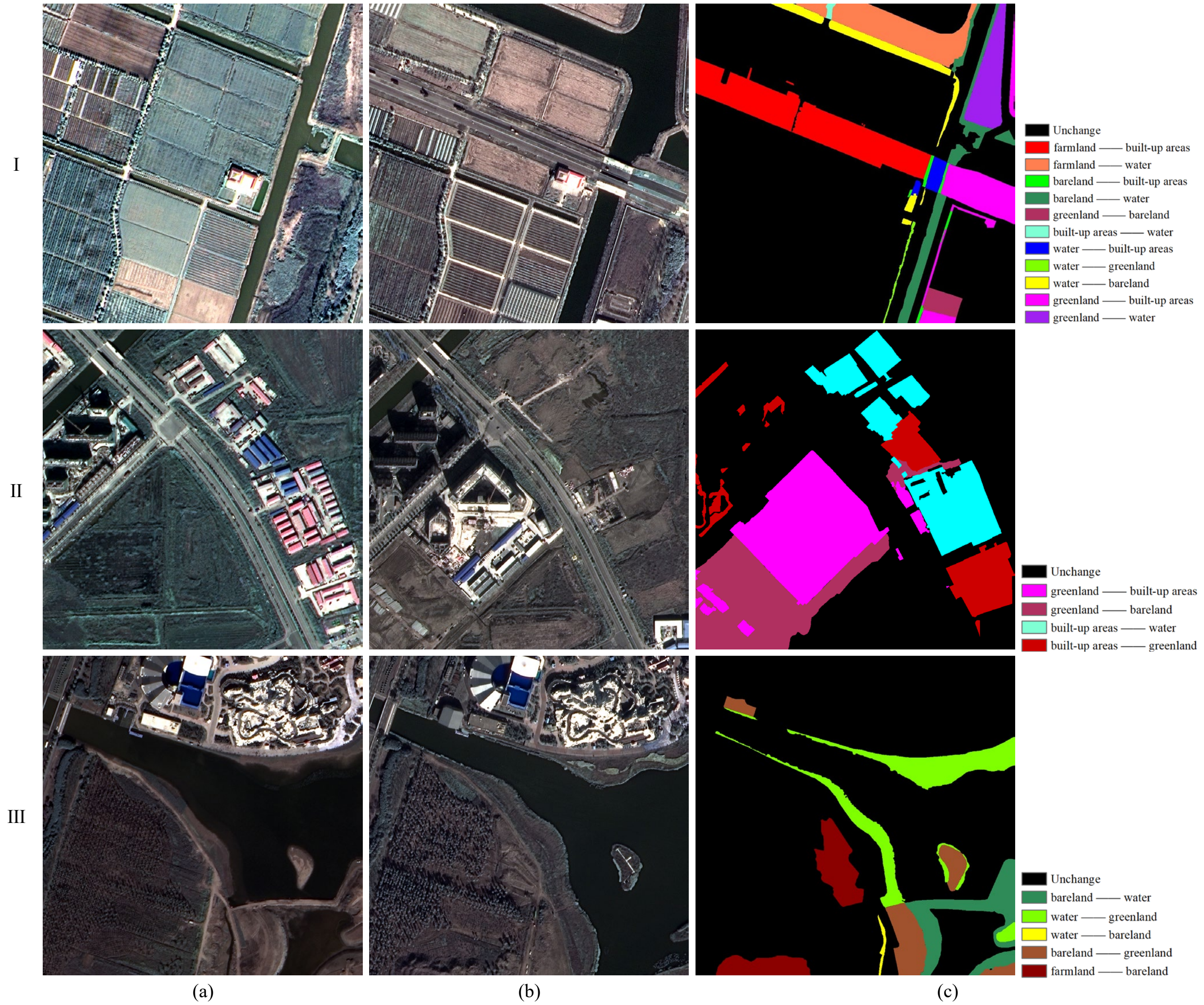


Fig. 5. True-color images and ground truth maps for the three datasets. (a) T1 image. (b) T2 image. (c) Ground truth (GT).

B. Benchmark methods

To fairly evaluate the proposed method, the DSCNH [9], DDSCN [44], CEGCN [36] and GAT-CNN fusion network (GCFN) were selected to conduct compared experiments.

DSCNH serves as a representative binary change detection method, incorporating multi-scale feature modules for change detection based on pixels. In this context, the model's output was modified to produce change category labels.

DDSCN is an end-to-end model designed for large-sample input, utilizing depth-separable convolution [45] and the U-Net [46] structure, and stands as a representative model for aerial image change detection. For model training, we extracted 1/4 of the regions from each dataset to construct the training data with a sample size of 112×112, while the remaining regions were utilized to create the test data. Ultimately, we obtained 1000 training image pairs, each sized 112×112.

CEGCN, initially designed for image classification, is introduced for feature propagation between image pixels and graph nodes, effectively addressing the challenge of data structure disparities between CNN and GCN, thus enabling them to collaborate seamlessly within a single network. To best of our understanding this model is firstly introduced into the domain of multi-class change detection with this study to integrate feature propagations between CNN and GAT. We have selected CEGCN as a comparative benchmark to validate the efficacy of incorporating the attention mechanism in this context.

Furthermore, we have replaced the P-GAT component of the method of this paper with the traditional Graph Attention Network (GAT) [43] to create GCFN serving as a second comparative test to discuss the influence of geospatial knowledge on deep learning models.

C. Accuracy evaluation metrics

In addition to the commonly used metrics like overall accuracy (OA), kappa coefficient (KPP) and average accuracy (AA), we introduce another additional metrics for MCD, i.e., Intersection over Union (IoU). Furthermore, we calculate the Mean Intersection over Union (MIoU), which is the average of individual IoU values.

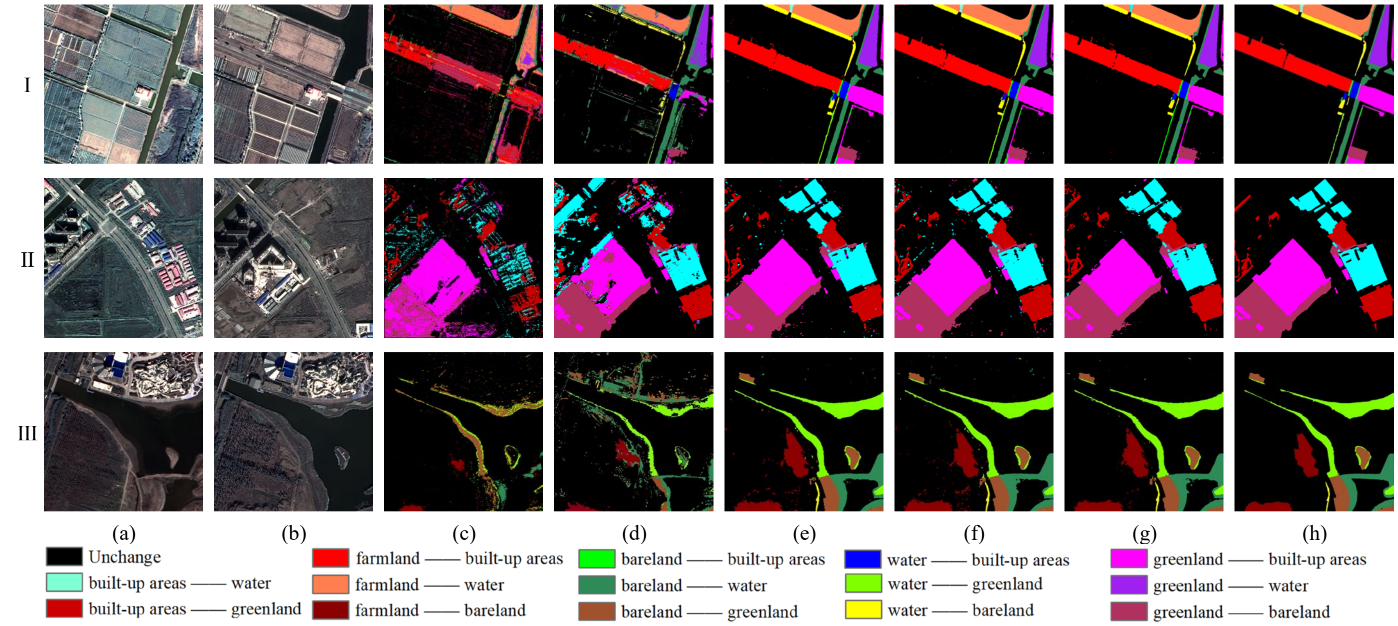


Fig. 6. Visualization of multi-class change detection results on three datasets. (a) T1 image. (b) T2 image. (c) DSCNH. (d) DDSCN. (e) CEGCN. (f) GCFN. (g) PGCFN. (h) Ground-truth (GT).

Thus, we employ a total of four accuracy evaluation metrics in this section. Their definitions are as follows:

$$OA = \frac{TP + PN}{TP + TN + FN + FP} \tag{16}$$

$$AA = sum(recall) / n \tag{17}$$

$$recall = \frac{TP}{TP + FN} \tag{18}$$

$$IOU = \frac{TP}{TP + FP + FN} \tag{19}$$

$$KPP = \frac{OA - Pe}{1 - Pe} \tag{20}$$

$$Pe = \frac{(TP + FP) \times (TP + FN) + (TN + FN) \times (TN + FP)}{(TP + TN + FP + FN)^2} \tag{21}$$

where TP indicates the number of positive examples classified accurately, FP is the number of actual negative examples classified as positive, FN means the number of actual positive examples classified as negative, and TN shows the number of negative examples classified accurately.

D. Implementation details

The segmentation scales are set as 800, 1500, and 1300 for the three datasets, respectively, and the number of components *l* is fixed to 25. For model training and validation, a random selection of 3% and 3% of samples per class is drawn from each of the three datasets. The experiments were conducted on an RTX-3080 GPU. We employed the Adam optimizer [47] with a learning rate of 0.0005 to train our network. Each experiment was repeated ten times, and the reported results represent the mean and standard deviation of each accuracy evaluation metric. It's important to note that all baseline methods were configured with hyperparameters as recommended in their respective original papers.

E. Experiments results

The result maps generated by these methods are illustrated in Fig. 6, and change detection accuracies are detailed in Tables I–III. In-depth analysis of the detection results obtained through different methods is shown in Figs. 7-9.

Fig. 6 illustrates the limitations of DSCNH when applied to high-resolution data. In this context, pixel-based change detection proves ineffective in learning meaningful information on our datasets. Conversely, DDSCN, employing Image-to-Image learning, surpasses the detection performance achieved through Point-to-Point learning (DSCNH). The other three methods exhibit a capacity to detect changes in urban areas and categorize them to some degree, highlighting the effectiveness of fusing pixel-level and object-level features in extracting valuable information from high-resolution images. Notably, the method proposed in this paper excels in producing superior results, particularly in capturing finer details with higher precision.

For Dataset I, the changed regions mainly comprise new constructions and high-complexity features, and contains a large number of changes in farmland and water body. The newly constructed planting greenhouse in the lower left corner is prone to being mistakenly detected as a change. Moreover, there is a serious imbalance of variation categories in the data set, our method surpasses all the compared methods. As is shown in Table I, the MIoU and Kappa of PGCFN are increased by 3.3% and 1.1% when compared with CEGCN, respectively, and by 1.5% and 1.08% compared with GCFN. And DDSCN produces less favorable results on Dataset I compared to PGCFN. It is evident from the detailed plots (as illustrated in Fig. 7) that the PGCFN model's detection results closely align with the labels, outperforming the other models. This is evident

TABLE I
ACCURACY ACHIEVED BY THE DIFFERENT MODELS ON DATASET I.

	Change type	DSCNH	DDSCN	CEGCN	GCFN	PGCFN
IOU	Farmland-Bulit up areas	28.03±1.78	22.98±0.25	94.19±0.09	**95.53±0.06**	95.24±0.03
	Farmland-Water	47.95±0.92	30.75±1.21	97.98±0.10	97.92±0.05	**98.14±0.07**
	Bareland-Bulit up areas	34.38±0.84	38.68±0.92	70.49±1.89	75.44±1.03	**85.84±0.17**
	Bareland-Water	30.79±1.63	30.16±0.45	90.07±0.26	89.48±0.16	**90.09±0.27**
	Water-Bulit up areas	15.35±1.01	28.59±1.09	93.21±0.34	93.68±0.48	**95.05±0.22**
	Water-Greenland	27.53±1.18	29.32±0.91	70.29±1.24	71.38±0.77	**72.05±0.54**
	Water-Bareland	26.13±1.26	45.15±0.71	90.99±0.13	89.74±0.16	**91.62±0.12**
	Greenland-Bulit up areas	25.47±0.92	32.34±0.64	91.52±0.09	91.17±0.14	**92.45±0.16**
	Greenland-Water	30.77±0.81	43.45±0.23	96.92±0.05	97.40±0.12	**97.42±0.14**
	Greenland-Bareland	14.48±2.65	51.16±0.52	83.18±0.31	90.09±0.26	**91.17±0.19**
	Bulit up areas-Greenland	13.64±2.70	54.21±0.32	76.85±2.46	**86.24±1.38**	84.12±1.32
	No change	81.43±0.62	84.51±0.16	98.08±0.04	98.18±0.02	**98.24±0.18**
MIoU		31.33±1.85	47.65±0.42	87.87±0.34	89.69±0.35	**91.20±0.12**
OA		54.68±0.34	75.60±0.37	98.34±0.04	98.39±0.01	**98.76±0.06**
KPP		12.74±0.92	38.50±0.48	95.98±0.09	96.01±0.04	**97.09±0.01**
AA		59.32±0.87	74.29±0.61	97.20±0.34	97.55±0.24	**98.21±0.12**

TABLE II
ACCURACY ACHIEVED BY THE DIFFERENT MODELS ON DATASET II.

	Change type	DSCNH	DDSCN	CEGCN	GCFN	PGCFN
IOU	Greenland-Bulit up areas	60.93±0.27	76.10±0.32	**96.54±0.08**	96.43±0.10	96.35±0.04
	Greenland-Bareland	43.61±0.38	69.67±0.12	94.08±0.07	93.74±0.13	**95.74±0.06**
	Bulit up areas-Greenland	41.79±0.29	39.89±0.36	85.80±0.11	87.34±0.13	**88.27±0.15**
	Bulit up areas-Bareland	26.64±0.96	34.89±0.26	91.07±0.07	89.95±0.16	**91.86±0.12**
	No change	79.43±0.16	70.80±0.09	95.74±0.18	95.55±0.28	**96.15±0.24**
MIoU		50.48±0.24	58.27±0.18	92.64±0.11	92.06±0.16	**94.07±0.12**
OA		79.22±0.33	77.52±0.16	97.12±0.12	96.89±0.17	**98.54±0.16**
KPP		60.02±0.81	65.15±0.20	95.06±0.20	94.68±0.28	**96.06±0.28**
AA		75.91±0.46	78.24±0.19	98.42±0.06	98.35±0.05	**98.48±0.08**

TABLE III
ACCURACY ACHIEVED BY THE DIFFERENT MODELS ON DATASET III.

	Change type	DSCNH	DDSCN	CEGCN	GCFN	PGCFN
IOU	Farmland-Bareland	23.24±0.31	40.10±0.45	87.60±0.26	86.90±0.37	**87.88±0.10**
	Bareland-Water	21.57±0.67	36.15±0.23	92.63±0.23	92.70±0.37	**93.38±0.14**
	Bareland-Greenland	30.14±0.28	45.67±0.12	**93.89±0.18**	90.13±0.14	91.58±0.15
	Water-Greenland	17.99±0.46	61.79±0.36	92.28±0.12	91.90±0.10	**92.52±0.13**
	Water-Bareland	18.17±0.64	46.03±0.30	78.36±0.37	76.94±0.70	**80.05±0.74**
	No change	86.01±0.23	90.60±0.09	98.21±0.16	98.01±0.20	**98.47±0.14**
MIoU		32.84±0.22	53.88±0.38	90.49±0.28	89.43±0.32	**91.90±0.22**
OA		83.72±0.19	89.00±0.52	98.29±0.13	98.15±0.16	**99.11±0.12**
KPP		25.19±0.47	65.05±0.31	94.29±0.41	93.86±0.53	**95.75±0.37**
AA		71.57±0.10	83.23±0.20	98.57±0.05	98.54±0.16	**98.70±0.02**

in its inference results, which closely match the ground truth, featuring fewer error pixels and smoother feature boundaries.

The PGCFN model shows higher accuracy in identifying small-scale features, highlighting its advantages in identifying tiny objects. In the change area of Dataset II, we observe the presence of both new buildings and demolished buildings. However, in the results obtained from CEGCN and GCFN for Dataset II, some noise is evident, as depicted in Figs. 6 II(d)-(e). Notably, in the west of the image, areas covered by building shadows are inaccurately identified as changes towards the water body. This is due to the higher spectral similarity between shadows and water. Despite these challenges, our model consistently outperformed others on Dataset II, as shown in Table II. Generally, other models often misclassified certain pseudo-changes arising from color variations as actual change areas, particularly in cases involving alterations in vegetation. These models exhibited lower performance in correctly identifying such changes.

Similarly, PGCFN achieved best results on Dataset III. The MIoU of PGCFN is superior to that of GCFN by over 2.47%, and the Kappa is increased by 1.46% when compared with CEGCN, as shown in Table III. Briefly, our model obtained results with more complete boundaries and fewer noise points. The pixel-based deep learning model DSCNH exhibited a high rate of false detections and missing detections in our dataset. This could be due to the high level of detail, numerous categories, and increased complexity present in data with a 0.5-meter resolution, which poses greater challenges for the model. DDSCN achieves better results compared to DSCNH, benefiting by the end-to-end full convolution structure. It is worth noting that GCFN is indeed a model combining CNN and GAT. The outcomes of this combination demonstrate that integrating CNN and GNNs can achieve good results. We hope these findings will inspire and guide future research efforts.

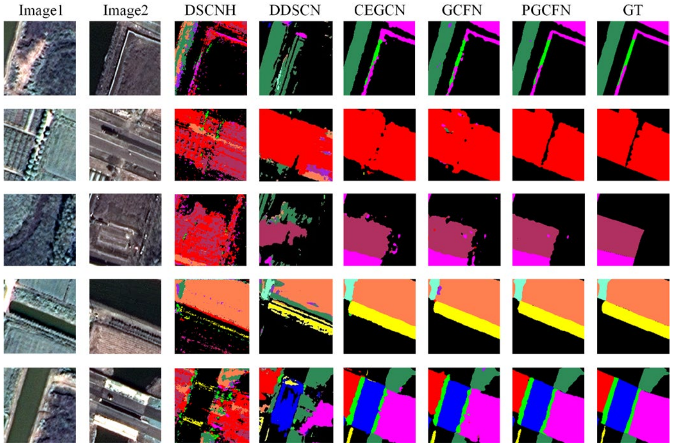


Fig. 7. Visualization of detail results on Dataset I.

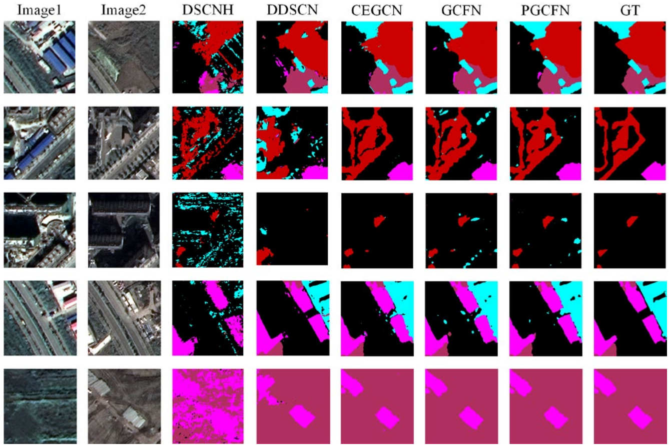


Fig. 8. Visualization of detail results on Dataset II.

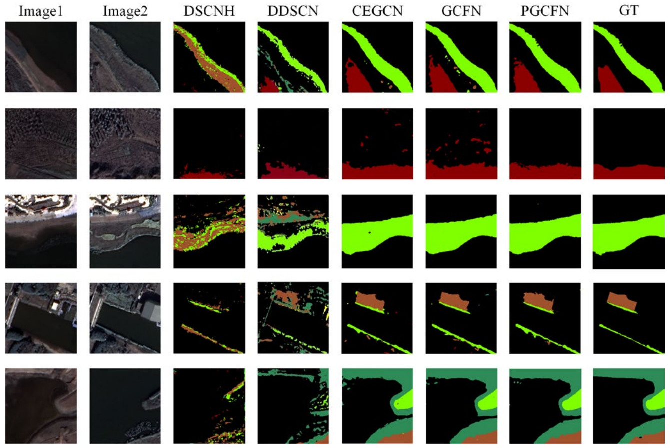


Fig. 9. Visualization of detail results on Dataset III.

V. Discussion

A. Influence of the component value l

The proposed method encodes spatial information using a one-hot encoder and generates a multi-dimensional relationship matrix by dividing the scalar distance into l components. We selected eight l values, specifically [5, 10, 15, 20, 25, 30, 35, 40, 45], to analyze the influence of different component values on the change detection results. The performance of the proposed model under different component values are illustrated in Fig. 10.

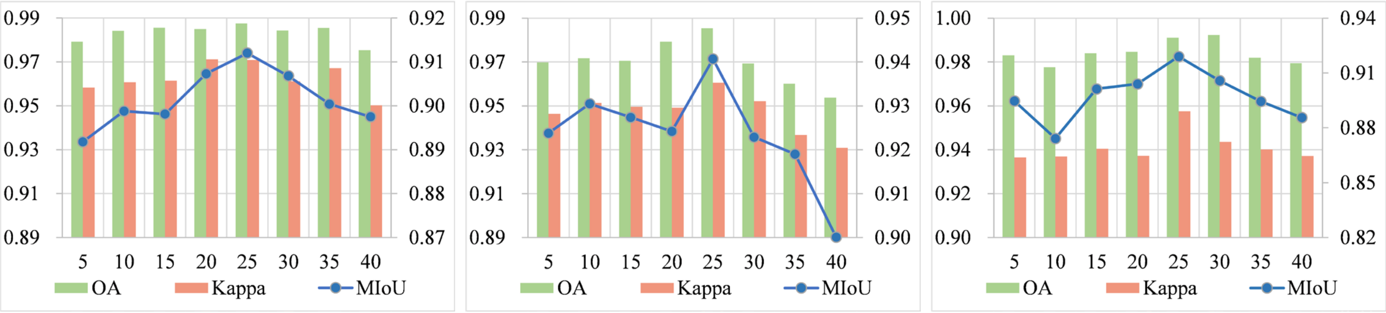
Based on change detection accuracy, the optimal performance is achieved when the component values for the three datasets are set to 25, as shown in Figs. 10(a)-(c). When the number of component values is small, such as 5 and 10, it can lead to lower change detection accuracy since fewer components fail to adequately capture the complex spatial information. Conversely, when the l-value is large, such as 40, it increases computational costs and doesn't necessarily result in improved change detection performance. This analysis helps determine the most appropriate component values to choose in a given situation to strike a balance between computational costs and performance requirements.

B. Visualization and analysis of attention mechanism

This study achieved a profound integration of geospatial theory and deep learning through the application of graph neural network. Graph structures were used to simulate the complex spatial relationships, encompassing aspects such as distance, as well as homogeneity and heterogeneity among ground objects. Beyond assessing the model's efficacy using detection accuracy, it is also valuable to qualitatively investigate the quality of the learned feature representations. To achieve this, we provided visualizations of the transformed feature representations obtained from the initial layer of the P-GAT.

In this context, the Python package Networkx was adopted to visualize subgraphs. Initially, we normalized the attention coefficients (i.e., weights) within the range of (1, 5]. We employed a tree-like graph layout to visually represent the nodes and their neighboring nodes. Specifically, the node is positioned at the center, with its neighbors arranged in a circular fashion. Node categories are represented using color coding, and the link size corresponds to the weight—thicker lines indicate higher weights. In addition, a multivariable line chart illustrates the change of node distances and attention weights. The distance variable is depicted in blue, whereas the weight variable is represented by the red line. It is important to note that the selected visualization results stem from three datasets, underscoring their generalizability.

Through an extensive series of experiments, we have observed that normal graph attention models yield features that exhibit minimal variations across different nodes. As depicted in Figs. 11(b-1), (d-1), and (e-1), the weights assigned to nodes tend to concentrate within a narrow range. This implies that distinct ground object types and various spatial positions exert similar influences on one another. Normal graph attention networks are typically utilized to uncover the relationships and interactions between nodes in graph data [48]. However, when dealing with complex data, these models often struggle to capture subtle distinctions among different nodes. Consequently, the learned weights exhibit similarities among various nodes, may leading to limitations in model performance [49].



(a) (b) (c)

Fig. 10. Detection accuracy under different component values. The left sub-axis represents the OA and Kappa measurements, while the right sub-axis represents the MIoU measurements, and the horizontal axis corresponds to the *l*-value. (a) Dataset I. (b) Dataset II. (c) Dataset III.

As illustrated in Figs. 11(a-4) to (e-4), the results obtained through P-GAT reveal a notable disparity in the weights assigned to its neighboring nodes. P-GAT incorporates geospatial information, enabling the model to place greater emphasis on geographic proximity when evaluating connections between nodes. This means that nodes in closer geographic proximity receive heightened geographic attention, enhancing the model's ability to capture geospatial patterns more effectively.

C. Computational efficiency and model complexity

Training and testing times, as well as model complexities for each method on every dataset, are detailed in Tables IV and V, respectively. The multi-class change detection task requires assigning a label to every pixel in a multi-temporal image, with the time required to classify all pixels considered as the test time. As indicated in Table IV, integrating CNN and GCN leads to faster convergence. In comparison to other methods, our approach integrates spatial principle of superpixels and utilizes the entire image as input, thereby significantly enhancing the efficiency of both training and testing.

TABLE IV
RUNNING TIME ON EACH DATASET.

Network	Time	Dataset I	Dataset II	Dataset III
DSCNH	Train	498.26s	481.98s	495.21s
	Test	120.42s	121.57s	120.61s
DDSCN	Train	1821.92s	1107.53s	1903.01s
	Test	6.39s	6.01s	5.78s
GCFN	Train	82.73s	81.09s	85.62s
	Test	2.3×10^{-2}s	1.9×10^{-2}s	2.1×10^{-2}s
CEGCN	Train	95.05s	93.41s	81.79s
	Test	4.9×10^{-2}s	4.3×10^{-2}s	5.0×10^{-2}s
PGFCN	Train	80.13s	72.41s	68.79s
	Test	0.83×10^{-2}s	0.77×10^{-2}s	0.90×10^{-2}s

As illustrated in Table V, we conducted a comprehensive comparison of the model complexity for each method. FLOPs, denoting the number of floating-point operations a computing entity can complete in one second, and the number of parameters (#Params), representing the count of learnable weights and biases in the model, can be combined to gauge the overall complexity of the network. Notably, DDSCN exhibited the highest parameter count, while PGFCN boasted a relatively modest number of parameters, closely aligned with GCFN. Despite DSCNH's lightweight nature, it falls short in performance on high-resolution data compared to other models. Analyzing the interplay between Params and FLOPs, PGCFN stands out with a notable volume of floating-point operations with relatively few parameters.

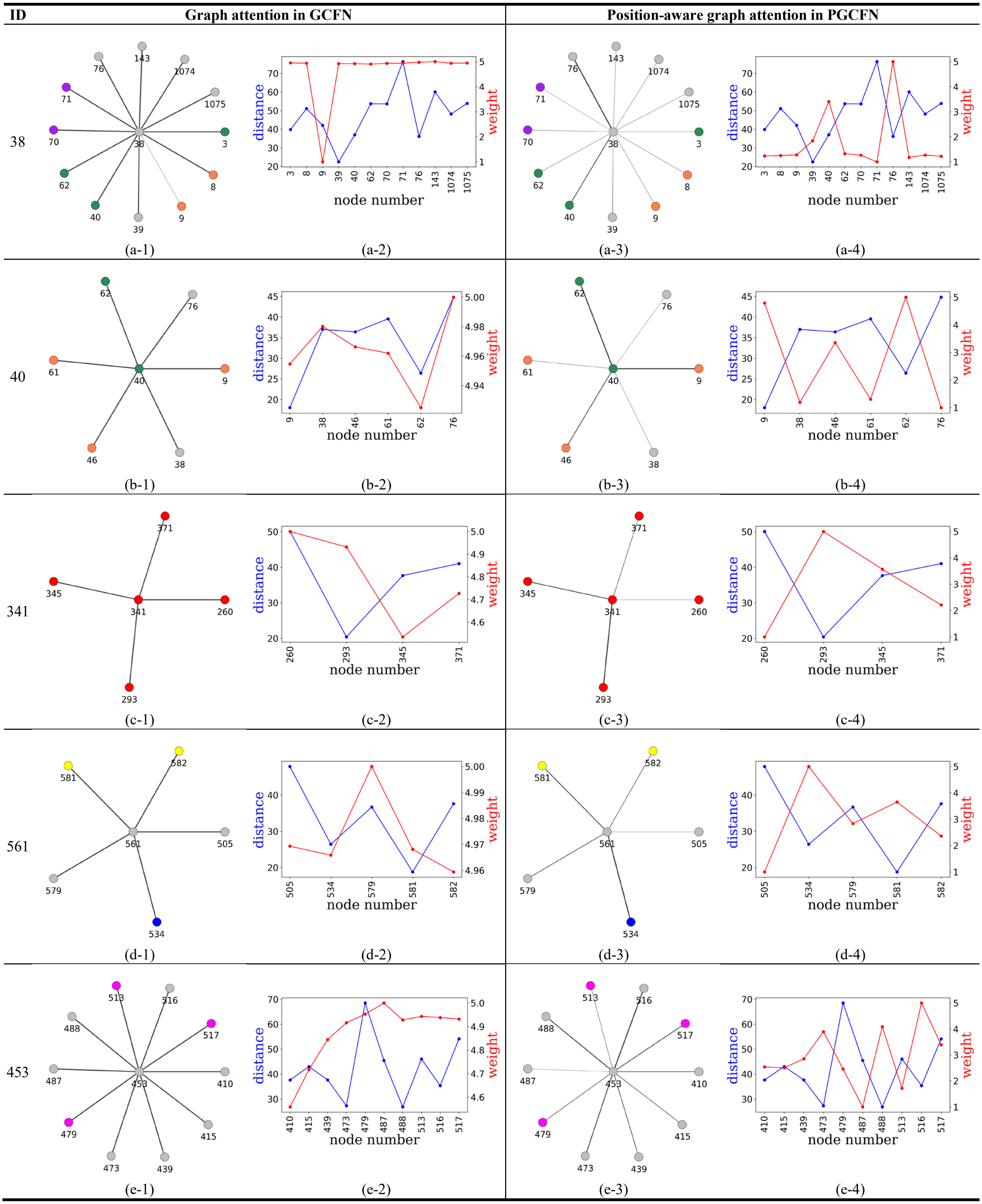
TABLE V
THE COMPARISON OF PARAMS AND FLOPS FOR NETWORKS.

Network	#Params	#FLOPs
DSCNH	5.58×10^{3}K	0.65G
DDSCN	29.82×10^{3}K	86.06G
GCFN	55.96K	56.29G
CEGCN	140.12K	49.22G
PGFCN	55.19K	55.52G

D. Robustness study

To quantitatively assess the robustness of our method, we chose common Salt & Pepper noise as well as Strip noise in remote sensing images to simulate noise and investigate the performance of various algorithms on our datasets. We introduced Salt & Pepper noise to bi-temporal images at varying noise rates ranging from 5% to 45% with 10% intervals, utilizing mIOU as the primary evaluation metric. As can be seen from Fig. 12 that PGCFN is more robust against noises compared to others.

Also, we introduced stripe noise with a varying noise rate ranging from 15% to 55% at intervals of 10% to assess the robustness of each algorithm. The results are presented in Fig. 13. It is evident that the model, integrating both CNN and GNN, excels in its ability to counter noise. This excellence can be attributed to the synergistic advantages of these two components in handling diverse data types and modeling intricate relationships. Our model also demonstrates heightened robustness, showcasing an enhanced capacity to handle complex data efficiently within a noisy environment.



(a-1) (a-2) (a-3) (a-4)

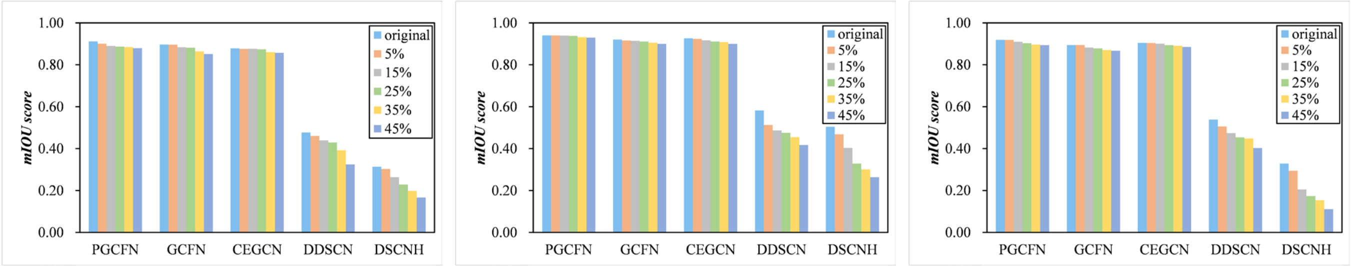
(b-1) (b-2) (b-3) (b-4)

(c-1) (c-2) (c-3) (c-4)

(d-1) (d-2) (d-3) (d-4)

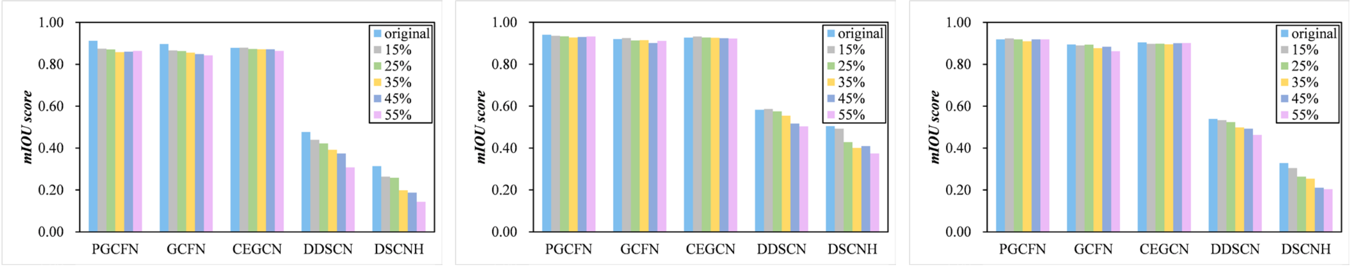
(e-1) (e-2) (e-3) (e-4)

Fig. 11. Examples of tree chart and multivariable line chart for different nodes. Where (*-1) is the tree charts on nodes generated by GCFN, (*-2) is the multivariable line charts of corresponding spatial distances and weights; (*-3) is the tree charts on nodes generated by PGCFN, and (*-4) is the multivariable line charts of corresponding spatial distances and weights.



(a) (b) (c)

Fig. 12. Robustness analysis using Salt & Pepper noise with the different noise rates on (a) Dataset I, (b) Dataset II, and (c) Dataset III.



(a) (b) (c)

Fig. 13. Robustness analysis using Stripe noise with the different noise rates on (a) Dataset I, (b) Dataset II, and (c) Dataset III.

VI. Conclusion

Deep learning models are characterized by a multitude of network parameters, high complexity, and often render the decisions and intermediate processes less interpretable. Current intelligent analysis methods for remote sensing images based on deep learning mainly focus on architectural innovations [50]. This study sought to bridge the gap by leveraging graph neural network theory to achieve a deep integration of geospatial knowledge and deep learning, and a multi-class change detection network named PGCFN was proposed. Building on neural network theory, we delved into the powerful synergies between deep learning and geospatial knowledge, with the aim of empowering intelligent interpretation with enhanced geographical context understanding. The results demonstrate that the position-aware graph attention-based change detection model can adeptly extract change features from the bi-temporal images. It is evident that the incorporation of geospatial information enhances the model's spatial comprehension, leading to an improved accuracy in change detection. Furthermore, this study certifies that the combination of pixel and object-level features ensures the high precision of detection results. While the proposed change detection method has shown satisfactory results, there are still limitations that require attention in future research.

1) Variations in the number of superpixels generated can undeniably impact detection accuracy and due to computational constraints, our research has yet to comprehensively investigate the correlation between segmentation parameters and outcomes. We aim to investigate this in our coming future research and delve into change detection results under diverse segmentation conditions.

2) Remote sensing images of large scenes often encompass a diverse range of ground objects, making changes within them more complex. Although the data-driven change detection model can achieve excellent results on small-scale data, their detection accuracy tends to be decreased when applied to large-scale data due to the influence of sample quality and quantity. However, the geographical spatial principles are not constrained by the scale of the scene. Building on the findings of this paper, the utilization of spatial principle will continue to be enriched in the future, enhancing the potential for large-scale image change detection within the spatial domain.

3) Another challenge relates to the limited samples in supervised learning methods. Supervised learning is particularly sensitive to sample acquisition difficulties and inadequate sample sizes. The issue of sample imbalance further compounds the challenges associated with supervised learning applications. Future studies will be directed to conduct a more extensive set of experiments in semi-supervised and unsupervised representation learning to address these issues.

Acknowledgment

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] W. Li and C.-Y. Hsu, "GeoAI for large-scale image analysis and machine vision: Recent progress of artificial intelligence in geography," *ISPRS International Journal of Geo-Information,* vol. 11, no. 7, pp. 385, 2022.

[2] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Multitask learning for large-scale semantic change detection," *Computer Vision and Image Understanding,* vol. 187, pp. 102783, 2019.

[3] K. Yang *et al.*, "Asymmetric siamese networks for semantic change detection in aerial images," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 60, pp. 1-18, 2021.

[4] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: a systematic survey," *IEEE transactions on image processing,* vol. 14, no. 3, pp. 294-307, 2005.

[5] P. R. Coppin and M. E. Bauer, "Digital change detection in forest ecosystems with remote sensing imagery," *Remote sensing reviews,* vol. 13, no. 3-4, pp. 207-234, 1996.

[6] A. Asokan and J. Anitha, "Change detection techniques for remote sensing applications: A survey," *Earth Science Informatics,* vol. 12, pp. 143-160, 2019.

[7] F. Luo, T. Zhou, J. Liu, T. Guo, X. Gong, and J. Ren, "Multiscale diff-changed feature fusion network for hyperspectral image change detection," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 61, pp. 1-13, 2023.

[8] W. Liu, Y. Lin, W. Liu, Y. Yu, and J. Li, "An attention-based multiscale transformer network for remote sensing image change detection," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 202, pp. 599-609, 2023.

[9] M. Wang, K. Tan, X. Jia, X. Wang, and Y. Chen, "A deep siamese network with hybrid convolutional feature extraction module for change detection based on multi-sensor remote sensing images," *Remote Sensing,* vol. 12, no. 2, pp. 205, 2020.

[10] S. Tian, Y. Zhong, Z. Zheng, A. Ma, X. Tan, and L. Zhang, "Large-scale deep learning based binary and semantic change detection in ultra high resolution remote sensing imagery: From benchmark datasets to urban application," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 193, pp. 164-186, 2022.

[11] H. Zhang, G. Ma, Y. Zhang, B. Wang, H. Li, and L. Fan, "MCHA-Net: A multi-end composite higher-order attention network guided with hierarchical supervised signal for high-resolution remote sensing image change detection," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 202, pp. 40-68, 2023.

[12] S. Saha, F. Bovolo, and L. Brurzone, "Unsupervised multiple-change detection in VHR optical images using deep features," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, 2018: IEEE, pp. 1902-1905.

[13] X. Wang *et al.*, "Double U-Net (W-Net): A change detection network with two heads for remote sensing imagery," *International Journal of Applied Earth Observation and Geoinformation,* vol. 122, pp. 103456, 2023.

[14] D. Wen *et al.*, "Change detection from very-high-spatial-resolution optical remote sensing images: Methods, applications, and future directions," *IEEE Geoscience and Remote Sensing Magazine,* vol. 9, no. 4, pp. 68-101, 2021.

[15] J. Hu and Y. Zhang, "Seasonal change of land-use/land-cover (LULC) detection using MODIS data in rapid urbanization regions: A case study of the pearl river delta region (China)," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* vol. 6, no. 4, pp. 1913-1920, 2013.

[16] D. Hong *et al.*, "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sensing of Environment,* vol. 299, pp. 113856, 2023.

[17] C. Li, B. Zhang, D. Hong, J. Yao, and J. Chanussot, "LRR-Net: An Interpretable Deep Unfolding Network for Hyperspectral Anomaly Detection," *IEEE Transactions on Geoscience and Remote Sensing,* 2023.

[18] X. Wu, D. Hong, and J. Chanussot, "UIU-Net: U-Net in U-Net for infrared small object detection," *IEEE Transactions on Image Processing,* vol. 32, pp. 364-376, 2022.

[19] L. Khelifi and M. Mignotte, "Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis," *Ieee Access,* vol. 8, pp. 126385-126400, 2020.

[20] Y. Zhou, J. Wang, J. Ding, B. Liu, N. Weng, and H. Xiao, "SIGNet: A Siamese Graph Convolutional Network for Multi-Class Urban Change Detection," *Remote Sensing,* vol. 15, no. 9, pp. 2464, 2023.

[21] Q. Zhu, X. Guo, Z. Li, and D. Li, "A review of multi-class change detection for satellite remote sensing imagery," *Geo-spatial Information Science,* pp. 1-15, 2022.

[22] I. Kotaridis and M. Lazaridou, "Remote sensing image segmentation advances: A meta-analysis," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 173, pp. 309-322, 2021.

[23] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral earth observation using convolutional neural networks," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, 2018: IEEE, pp. 2115-2118.

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

[25] L. Mou, L. Bruzzone, and X. X. Zhu, "Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 57, no. 2, pp. 924-935, 2018.

[26] L. Ding, H. Guo, S. Liu, L. Mou, J. Zhang, and L. Bruzzone, "Bi-temporal semantic reasoning for the semantic change detection in HR remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 60, pp. 1-14, 2022.

[27] D. Wang, F. Zhao, C. Wang, H. Wang, F. Zheng, and X. Chen, "Y-Net: A multiclass change detection network for bi-temporal remote sensing images," *International Journal of Remote Sensing,* vol. 43, no. 2, pp. 565-592, 2022.

[28] L. Wang, J. Yan, L. Mu, and L. Huang, "Knowledge discovery from remote sensing images: A review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* vol. 10, no. 5, pp. e1371, 2020.

[29] Y. Yao, D. Suonan, and J. Zhang, "Compilation of 1: 50,000 vegetation type map with remote sensing images based on mountain altitudinal belts of Taibai Mountain in the North-South transitional zone of China," *Journal of Geographical Sciences,* vol. 30, pp. 267-280, 2020.

[30] W. Li, C.-Y. Hsu, and M. Hu, "Tobler's First Law in GeoAI: A spatially explicit deep learning model for terrain feature detection under weak supervision," *Annals of the American Association of Geographers,* vol. 111, no. 7, pp. 1887-1905, 2021.

[31] Y. Li, S. Ouyang, and Y. Zhang, "Collaboratively boosting data-driven deep learning and knowledge-guided ontological reasoning for semantic segmentation of remote sensing imagery," *arXiv preprint arXiv:2010.02451,* 2020.

[32] D. Hong *et al.*, "Spectralgpt: Spectral foundation model," *arXiv preprint arXiv:2311.07113,* 2023.

[33] H. Lyu, H. Lu, and L. Mou, "Learning a transferable change rule from a recurrent neural network for land cover change detection," *Remote Sensing,* vol. 8, no. 6, pp. 506, 2016.

[34] Y. Li, R. Chen, Y. Zhang, M. Zhang, and L. Chen, "Multi-label remote sensing image scene classification by combining a convolutional neural network and a graph neural network," *Remote Sensing,* vol. 12, no. 23, p. 4003, 2020.

[35] S. Saha, L. Mou, X. X. Zhu, F. Bovolo, and L. Bruzzone, "Semisupervised change detection using graph convolutional network," *IEEE Geoscience and Remote Sensing Letters,* vol. 18, no. 4, pp. 607-611, 2020.

[36] Q. Liu, L. Xiao, J. Yang, and Z. Wei, "CNN-enhanced graph convolutional network with pixel-and superpixel-level feature fusion for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 59, no. 10, pp. 8657-8671, 2020.

[37] W. R. Tobler, "A computer movie simulating urban growth in the Detroit region," *Economic geography,* vol. 46, no. sup1, pp. 234-240, 1970.

[38] Y. Ge, X. Zhang, P. M. Atkinson, A. Stein, and L. Li, "Geoscience-aware deep learning: A new paradigm for remote sensing," *Science of Remote Sensing,* vol. 5, pp. 100047, 2022.

[39] J. Hagenauer and M. Helbich, "A geographically weighted artificial neural network," *International Journal of Geographical Information Science,* vol. 36, no. 2, pp. 215-235, 2022.

[40] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence,* vol. 34, no. 11, pp. 2274-2282, 2012.

[41] R. Ji, K. Tan, X. Wang, C. Pan, and L. Xin, "PASSNet: A Spatial-Spectral Feature Extraction Network with Patch Attention Module for Hyperspectral Image Classification," *IEEE Geoscience and Remote Sensing Letters,* 2023.

[42] S. Ghaffarian, J. Valente, M. Van Der Voort, and B. Tekinerdogan, "Effect of attention mechanism in deep learning-based remote sensing image processing: A systematic literature review," *Remote Sensing,* vol. 13, no. 15, pp. 2965, 2021.

[43] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903,* 2017.

[44] R. Liu, D. Jiang, L. Zhang, and Z. Zhang, "Deep depthwise separable convolutional network for change detection in optical aerial images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* vol. 13, pp. 1109-1118, 2020.

[45] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251-1258.

[46] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 2015: Springer, pp. 234-241.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980,* 2014.

[48] G. Brauwers and F. Frasincar, "A general survey on attention mechanisms in deep learning," *IEEE Transactions on Knowledge and Data Engineering,* 2021.

[49] J. B. Lee, R. A. Rossi, S. Kim, N. K. Ahmed, and E. Koh, "Attention models in graphs: A survey," *ACM Transactions on Knowledge Discovery from Data (TKDD),* vol. 13, no. 6, pp. 1-25, 2019.

[50] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artificial intelligence review,* vol. 53, pp. 5455-5516, 2020.