Cross-Perception and Hierarchical Similarity Metric Network for Remote Sensing Image Change Detection

Haicheng Qu, *Member, IEEE*, and Lijuan Zhang

***Abstract*—Change detection (CD) is a fundamental operation in remote sensing image interpretation. This process employs a range of image processing and recognition techniques to identify semantic alterations in the same geographical region across different temporal phases. However, most of existing CD methods rarely utilize the relationship between dual-time phase features. In addition, they tend to overlook the potential benefits of integrating spatial and channel information, which impairs their ability to discern fine details and address pseudochanges. To address these limitations, we propose a cross-perception and hierarchical similarity metric network (CPHSM-Net). The features are first captured using a feature extractor that adds an adaptive spatial channel enhancement (ASCE) strategy to adaptively obtain a more meaningful representation of the features. Then, the relationships between the features of each layer are captured by the cross-perception (CP) module. Finally, the variation feature description is further enhanced by the hierarchical similarity metric (HSM) module, which is designed to capture the variations and differences in the images. The F1 scores obtained by CPHSM-Net in experimental tests on three publicly available datasets (LEVIR-CD, WHU-CD, and DSIFN-CD) were 91.07%, 92.56%, and 65.75%, respectively, which is superior to the state-of-the-art comparison methods.**

***Index Terms*—Adaptive spatial channel enhancement (ASCE), change detection (CD), cross-perception (CP), hierarchical similarity metric (HSM), remote sensing images, Siamese network.**

I. Introduction

CHANGE detection (CD) obtains relevant object changes by comparing dual-time-phase remote sensing images [1], [2]. CD plays a fundamental role in urban expansion research [3], land resource management [4], [5], and environmental surveillance [6]. With the maturing of remote sensing imaging technology and the great improvement in computing power, the information in remote sensing images is becoming more abundant, and thus the task of CD also faces great challenges.

The key challenge in CD is the extraction of "semantic changes" while suppressing "nonsemantic changes" "Semantic changes" are often the appearance or disappearance of objects that are specifically applied or defined, such as buildings and vegetation. "Nonsemantic changes" typically include shadows caused by light angles and seasonal changes in vegetation, which are also known as "pseudochanges." In the realm of traditional CD techniques, the preponderance relies on either manual extraction of features [7] or the employment of predefined thresholds for demarcating changes, such as change vector analysis [8], Gabor filter [9] and principal component analysis [10]. They are susceptible to the influence of "pseudochange" and therefore cannot effectively extract the change features. Concurrently, these methods are frequently intricate, necessitating a considerable number of computations, and are consequently susceptible to low robustness.

In the last few years, deep learning based methods have performed more prominently when performing feature extraction on complex images. These methods are often capable of deep feature representations such as convolutional neural networks (CNNs) [11], transformer [12]. Methods based on CNNs enable the network to learn data features in parallel through a special structure of shared local weights [13]. The transformer's remarkable capacity to model global dependencies enables it to mitigate the detriment of losing remote information [14]. Deep learning-based methods have a promising future in the much-anticipated field of hyperspectral image classification. In 2014, a multilayer SAE [15] was employed to extract both deep spectral and spatial features simultaneously. This is the first instance in which deep learning concepts have been applied to a hyperspectral image classification task. In order to overcome the shortcomings of graph neural network such as time-consuming and poor robustness, MRGAT [16] performs feature extraction by multiscale receptive field GAT. CMC-GAN [17] co-optimizes two groups of generative adversarial networks (GANs) to generate diversified scale samples. Domain adaptive methods have become a research hotspot for hyperspectral image classification. CCGDA [18] focuses on the class alignment problem in different domains to make the discriminative boundaries of the classifiers better adapt to the target domain.

Similarly, most deep learning methods propose powerful models to solve the CD problem but many of them still have some shortcomings. For example, FCN [19] uses the U-net model to roughly detect regions of new buildings, but it is less robust due to ignoring spatial information. PGA-SiamNet [20] uses a pyramid-based attention module to enhance the efficacy of CD,

Manuscript received 3 June 2024; revised 13 July 2024; accepted 30 July 2024. Date of publication 5 August 2024; date of current version 15 August 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 42271409 and in part by the Scientific Research Foundation of the Higher Education Institutions of Liaoning Province under Grant LJKMZ20220699. *(Corresponding author: Haicheng Qu.)*

The authors are with the Liaoning Technical University, Huludao 125105, China (e-mail: quhaicheng@lntu.edu.cn; ulrica_zhang@126.com).

Digital Object Identifier 10.1109/JSTARS.2024.3438246

but only uses differences or crosstabs to measure the similarity of images. To enhance the focus on salient change features while disregarding spurious "pseudochanges," DTCDSCN [21] focuses on correlations within the feature map, and DDCNN [22] and DSIFN [23] use spatial attention and channel attention to improve discrimination. These methods focus more on the feature map interior and emphasize the per-pixel channel importance. lightCDNet [24] employs an early fusion approach to enhance the feature extraction capability, which focuses on the channel information while processing the feature. ChangeFormer proposed by Bandara and Patel [25] employed a hierarchical inverse encoder to capture detailed and overall image features, but does not fully utilize crucial edge localization information.

Existing CD algorithms can detect change features well, but rarely correlate dual temporal phase features, while lacking the synergistic use of spatial and channel information. Based on this, this article proposes a CP and hierarchical similarity metric (HSM) network.

1) A new remote sensing image CD model is proposed, which includes an adaptive spatial channel enhancement (ASCE) strategy, a CP module and a HSM module. It can obtain the differences and correlations between dual-time-phase features, make good use of semantic information, and effectively eliminate "pseudochanges" to achieve better CD results.
2) The ASCE strategy is applied when extracting features to obtain a richer feature representation, realizing the synergistic use of channel information and spatial information without increasing the complexity of feature extraction.
3) The CP module enhances the interaction of feature information by using the correlation between and within the changing features to improve the model's recognition ability.
4) The HSM module divides the features into two hierarchical levels using the residual fusion method, and then uses the spatial information to measure the correlation between the features to improve detection performance.

The rest of this article is organized as follows. Section II reviews related work. Section III describes the proposed method in detail. To evaluate our approach, experiments are designed and the proposed method is discussed in Section IV. Finally, Section V concludes this article.

II. Related Work

A. Deep Learning Based Methods for CD

In recent years, Big Data analytics and AI have propelled CD to unprecedented heights, revolutionizing data processing, and utilization [26]. Numerous deep learning architectures, predominantly CNNs, have exhibited their prowess in the realm of CD. In the CD task, the model inputs two images at different times. A good model should be capable of processing the two input images in an efficient manner in order to generate accurate change maps. The seminal work on applying CNNs to the CD task was that of [27], where the authors proposed two methods for processing the input images. Many subsequent models have been developed based on these two approaches. The first approach is to process the two images separately using a Siamese network, with subsequent fusion of the features using different methods. The second approach is an early fusion strategy, whereby the two inputs are initially fused, and then subjected to subsequent operations such as feature extraction. To get the most out of images, deeper CNNs like ResNet [28] or VGG16 [29] are used to extract features. FC-Siam-diff [27] employs a symmetric network to extract two temporal features and subtracts them to obtain a change map. However, spectral and positional errors cause many false positives for the difference change map. To capture the information in the integrated features, a cross-stage combinatorial network was designed by Zhao et al. [30]. CLNet [31] improves the network's ability to extract more advanced semantic information about an image by creating a multistage fusion module. However, due to the fact that the sensory field of convolutional networks is fixed and the Siamese networks share weights, this limitation may lead to a weaker feature extraction capability of the model.

B. Attention Mechanisms

In recent years, a novel approach to identifying significant differences between spatial and channel features in computer vision tasks has emerged: the attentional mechanism [32]. It is capable of overcoming the inherent limitations of convolutional networks. The different channels correspond to different features. The utilization of the channel attention mechanism permits the network to dynamically identify and prioritize pivotal semantic data, thus enabling an adaptive selection of crucial information. In view of the above, SENet [33] was proposed as a means of collecting information from different channels and establishing relationships. Nevertheless, SENet acquires global information through global average pooling, which is unable to adequately model intricate target features, significantly constraining its modeling capabilities. Moreover, the process of establishing connections across diverse channels is inherently intricate and complex. To tackle the aforementioned challenges, Yang et al. [34] implemented a gated channel transformation approach for modeling the intricate relationships between various channels. On the other hand, the spatial attention mechanism amplifies the significance of the network's spatial location within a singular channel. To provide the network with the ability to detect the focal points of attention, the RAM model [35] is proposed. When extracting features, the STANet model [36] is designed with two different attentional mechanisms by exploiting the correlation of images from different temporal instances and spatial locations. This approach is designed to reduce the occurrence of false alarms in changing maps. However, their efficacy is reduced when applied to regions exhibiting minimal colour changes. The convolutional block attention module (CBAM) [37] includes a spatial awareness map in conjunction with a channel-specific attention map, thereby exploiting both spatial and channel-specific attention mechanisms. Despite this, CBAM sequentially cascades the spatial information and the channel information, which may result in the underutilization of spatial and channel information. The aforementioned attention-based CD methods have the potential to enhance CD

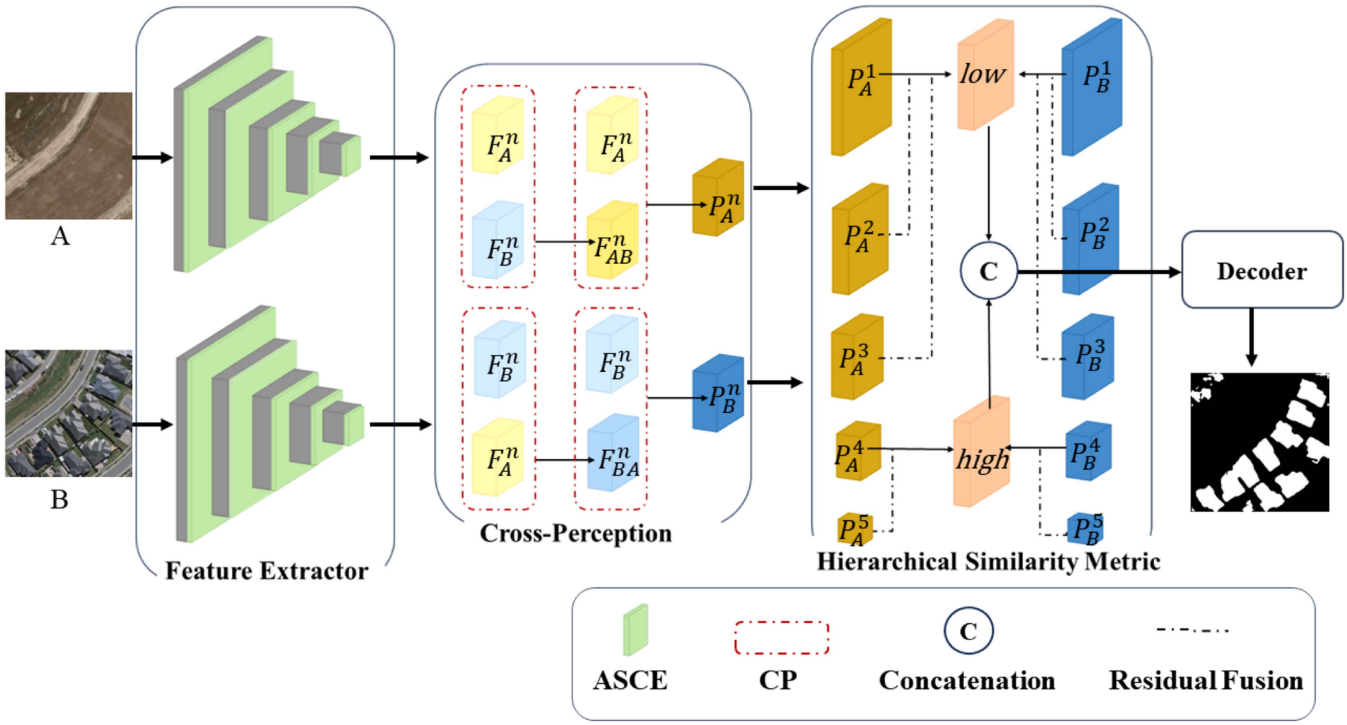

Fig. 1. CPHSM-Net structure.

performance. However, these methods do increase the model complexity to a certain extent, and at the same time, they are unable to effectively utilize the spatial and channel information in a synergistic manner.

C. Transformer

The transformer [38] has attained noteworthy achievements in the realm of computer vision and has furthermore been implemented for the CD task, serving as a global attention mechanism. The Vision Transformer (ViT) method [39] splits the input into homogeneous patches, and then transforms each of these patches into a normalised token of fixed length. This approach results in the generation of more discriminative features, although it is highly time-consuming. To address the inefficiency, Swin Transformer [40] employs a smaller window and patch interaction mechanism, thereby achieving a more optimal speed-accuracy tradeoff. The InterFormer [24] applied this attention mechanism to design IAM and combined it with a Siamese network for feature extraction effectively solving the problem that a Siamese network does not interact directly. In this article, we will apply this concept to the CP of features, the detection of subtle changes and the suppression of pseudochanges.

III. Methodology

A. Overview

In this section, we present an overview of the CPHSM-Net architecture and elaborate on the modules that constitute its core. As depicted in Fig. 1, the proposed CPHSM-Net comprises four principal components: a feature extractor, the CP module, the HSM module, and the decoder. Bitemporal images are used as inputs, and the more important feature representations are first obtained by feature adaptation with the addition of the ASCE strategy. To merge the feature mappings, the CP module is then applied at every scale to discern the relationship between the bi-chronological phase features. Then, different scales of the features are categorized into high and low level features to capture more spatial information using similarity metrics to capture changes and differences in the images. Finally, the decoder combines the results of similarity metrics to obtain the change map. Within the decoder, the feature map undergoes an upsampling process, employing a single interconnected state, along with two convolutional layers and two pixel reshuffling operations.

B. Feature Extractor

CPHSM-Net employs a CNNs backbone as its primary feature extractor. The feature extractor is divided into five layers, the first layer includes a 7×7 convolutional layer and an ASCE, and the remaining four layers are a residual block and an ASCE module, in which the residual block comes modified from Resnet-50 [41]. Commencing with a stride-2 convolutional layer for shallow feature extraction, followed by a stride-2 max pooling layer to decrease parameters, and integrating the ASCE module for channel and spatial information capture. In the next four-layer feature extractor, the residual block structure is based on the 1–4 layer structure in Resnet-50, where the feature mapping channels are reduced to 64/64/128/256/512. the features after each layer of residual block extraction are then utilized by further channel

information and spatial information extraction through ASCE. The specific ASCE strategy is detailed in the following section.

C. Adaptive Spatial Channel Enhancement Strategy

The residual block in ResNet-50 addresses gradient vanishing and emphasises overall image semantics, but may result in a limited focus on spatial and channel details in feature extraction. Therefore, we propose the ASCE strategy to obtain richer features during feature extraction. In Section II, we mentioned that the attention mechanism in deep learning is similar to that in living organisms, which can help the model to focus on more important information. There are three important concepts: query (Q), key (K), and value (V), which determines the importance of V by calculating the match between Q and K. The Nadaraya–Watson regression [42] uses a Gaussian kernel, which can be used as a nonparametric attention mechanism, and the formula is expressed as follows:

$$\mathbf{F} = \sum_{i=1}^{n} \frac{\exp\left(-\frac{1}{2}\left(\mathbf{Q}-\mathbf{K_i}\right)^2\right)}{\sum_{j=1}^{n}\exp\left(-\frac{1}{2}\left(\mathbf{Q}-\mathbf{K_j}\right)^2\right)}\mathbf{V_i} = \sum_{i=1}^{n}\mathrm{softmax}\left(-\frac{1}{2}\left(\mathbf{Q}-\mathbf{K_i}\right)^2\right)\mathbf{V_i} \tag{1}$$

where $\mathbf{F}$ denotes the feature map adapted by the attention mechanism, n denotes the number of feature vector dimensions, and $\mathbf{V}$, $\mathbf{Q}$, and $\mathbf{K}$ are taken from the feature map. The closer $\mathbf{K}$ gets to a given $\mathbf{Q}$, the more weight is given to $\mathbf{V}$, i.e., that $\mathbf{V}$ is said to "get more attention." Therefore, we can extend this to spatial and channel dimensions, and the ASCE strategy using (1) can be obtained

$$\mathbf{F} = \mathrm{Sigmoid}\left(\frac{(\mathbf{Q}-\mathbf{K})^2}{2\sigma^2}+\frac{1}{2}\right)\times\mathbf{V}. \tag{2}$$

In practice, the input feature map $\mathbf{X}$ as $\mathbf{K}$ and $\mathbf{V}$, $\mathbf{Q}$ is the average value in each channel dimension. σ^2 is obtained by calculating the spatial variance of each channel, and the larger value means that the more spatial information the channel contains. The attentional weights are adjusted by the spatial information on each channel so that the weights better represent the differences in spatial information between different channels. Often, a constant term is added to the original formula to provide a positive moderating effect of attentional weights $\mathbf{W}$. We introduce a constant 0.5, normalize it with a sigmoid, derive the attention weight, and multiply it with $\mathbf{V}$ to get a processed feature map $\mathbf{F}$. The detailed operation is shown in Fig. 2.

D. Cross-Perception Module

Suppose A and B are two input images, each with five layers of features reported after feature extraction, and $\mathbf{F}_A^n$ and $\mathbf{F}_B^n$ are used to denote the features extracted from inputs A and B at the nth layer, respectively. The conventional Siamese network operates by independently extracting features from the two inputs, without facilitating direct interaction between the extracted features. The CP module is designed to perceive the association and difference between the original features of

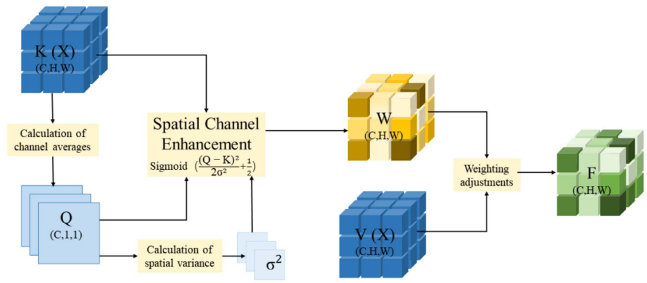

Fig. 2. ASCE strategy flowchart.

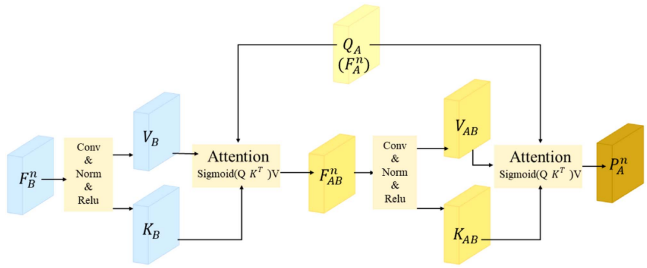

Fig. 3. CP module flowchart.

the features to enhance CD recognition. The detailed operation is shown in Fig. 3. Input $\mathbf{F}_A^n$ is used as $\mathbf{Q}_A$, input $\mathbf{F}_B^n$ is convolved operation to generate $\mathbf{K}_B$ and $\mathbf{V}_B$, then attention weights are generated by the dot product of $\mathbf{Q}_A$ and $\mathbf{K}_B$ and multiplied by $\mathbf{V}_B$ to retrieve the attention information. The detailed formula is the following:

$$\mathrm{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{softmax}(\mathbf{Q}\mathbf{K}^T)\mathbf{V}. \tag{3}$$

Connect the retrieved information in $\mathbf{F}_B^n$ with $\mathbf{F}_A^n$ to get the new feature $\mathbf{F}_{AB}^n$, which indicates that more attention is paid to information related to $\mathbf{F}_B^n$ in $\mathbf{F}_A^n$. Next, the same operation is performed for $\mathbf{F}_A^n$ with $\mathbf{F}_{AB}^n$, and finally, the cross-perceived feature $\mathbf{P}_A^n$ is obtained, associating the dual time-phase features for information exchange and enhancement. Similarly, $\mathbf{P}_B^n$ can be obtained by this module.

E. Hierarchical Similarity Metric Module

To further fuse and model multiscale features from feature extractors, CPHSM-Net uses this module to further exploit spatial information on multiscale features to measure their similarity. Deep features often contain global contextual information, which is useful for localising salient regions; shallow features contain details of spatial structure, which is useful for localizing boundaries [43]. Inspired by the above, we categorize the five layers of features into high and low level features through the residual fusion method. The residual fusion method is shown in Fig. 4. In deep networks, simple upsampling fusion may overwhelm small target features, so the current features are first deconvolutionalized to increase their dimensions, and then concatenated with the features of the previous layer. The concatenated features are pooled using a global average pooling operation to preserve important information. Finally they are

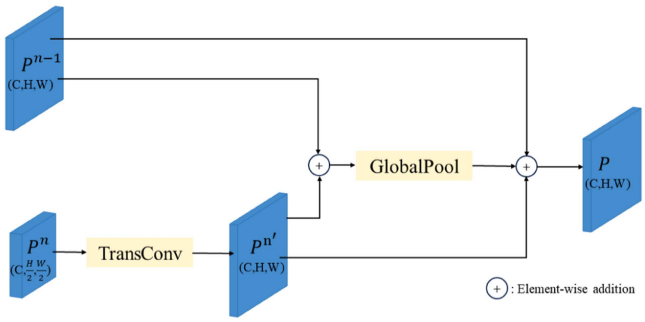

Fig. 4. Residual fusion flowchart.

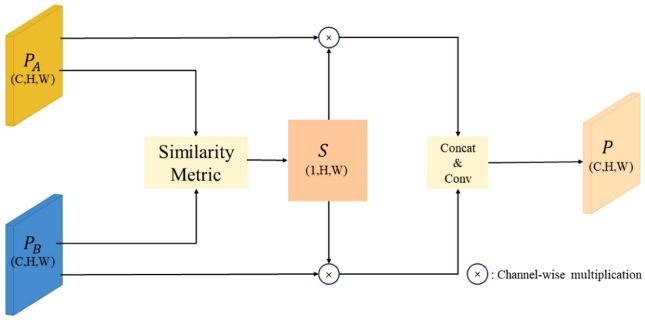

Fig. 5. Similarity metric flowchart.

combined using residual concatenation to obtain the fused features. High-level features are composed of extracted features from levels 3, 4, and 5, while low-level features are formed by combining features from levels 1 and 2.

The similarity metric is performed on the high-level features and low-level features, respectively. According to the characteristics of remote sensing images, the cosine-based similarity metric is used in the calculation. The results are calculated using the following formulas:

$$M(i,j) = \sum_{k=1}^{C} \mathbf{P}_{\mathrm{A}}^{k}(i,j) \times \mathbf{P}_{\mathrm{B}}^{k}(i,j) \tag{4}$$

$$N(i,j) = \sqrt{\sum_{k=1}^{C} (\mathbf{P}_{\mathrm{A}}^{k}(i,j))^2} \times \sqrt{\sum_{k=1}^{C} (\mathbf{P}_{\mathrm{B}}^{k}(i,j))^2} \tag{5}$$

$$\mathrm{S}(i,j) = 1 - \frac{M}{N}. \tag{6}$$

Here, C represents the channel count of the feature map, while $\mathbf{P}_A^k$ and $\mathbf{P}_B^k$ refer to the prechange and postchange feature maps of the input image in channel k, respectively. In addition, *i* and *j* denote the spatial coordinates and the corresponding similarity value within the feature map. Once S is acquired, $\mathbf{P}_A^k$ and $\mathbf{P}_B^k$ are individually multiplied, and subsequently, the resulting weighted feature maps are concatenated. This concatenation is then processed by a convolutional filter to incorporate layer-specific information. Finally the output of the features processed by this module $\mathbf{P}$. The similarity metric flowchart is shown in Fig. 5.

F. Loss Function

The cross-entropy loss function plays a crucial role in fine-tuning parameters and reducing loss metrics throughout the network's training phase. It serves as a tool to gauge the variance between the model's forecasts and the actual labels

$$\mathrm{Loss} = \frac{1}{N} \sum_{n=1}^{N} (-Y_n \log(P_n) - (1 - Y_n) \log(1 - P_n)). \tag{7}$$

The value of P_n represents the network result, while the value of Y_n represents the pixel status. Y_n is set to 0 to indicate that the pixel is unchanged, and Y_n is set to 1 to indicate that the pixel has been changed. Finally, the total number of pixels is represented by the value *N*.

IV. Experiments

A. Description of the Dataset

1) LEVIR-CD [44] is a large public and relatively clear (0.5 m/pixel) remote sensing dataset, containing 637 high-precision 1024×1024 pixel building images. The time span of the images is 5–14 years and includes minor changes in the buildings. Due to hardware limitations, the 1024×1024 pixel images are cropped into 16 nonoverlapping 256×256 pixel images, and the resulting 7120/1024/2048 pairs of samples and labels are used as the training, validation, and test sets, respectively.
2) WHU-CD [45] is a large-scale high-resolution remote sensing image dataset, containing a pair of high-resolution 32507×15354 (0.2 m/pixel) aerial images. It records changes in the New Zealand city of Christchurch before and after the earthquake, mainly building changes. They were cropped into mutually nonoverlapping images of 256 × 256 pixels and further filtered to remove too many images that did not make a difference. After filtering, the resulting 5928/480/552 pairs of samples and labels are used as the training, validation and test sets, respectively.
3) DSIFN-CD [23] contains high-resolution satellite images (2 m/pixel) of specific ground conditions in major cities in China, covering many ground variations. For each pair of 512 × 512 pixel images, they are sliced into nonoverlapping 256 × 256 pixel images, and the resulting 14400/1360/192 pairs of samples and labels are used as the training, validation and test sets, respectively.

B. Evaluation Metrics

In this article, precision (Pre), recall (Rec), F1-score (F1), and IoU are used as algorithm evaluation metrics to assess the effectiveness of the proposed algorithm. In the CD task, a higher Pre value indicates a reduced number of false positives in the prediction results, while a higher Rec value indicates a reduced number of false negatives. The overall performance of the prediction outcomes is evaluated using the F1-score and IoU metrics. As these values increase, the prediction results are deemed to be more accurate and superior. The formulae for these

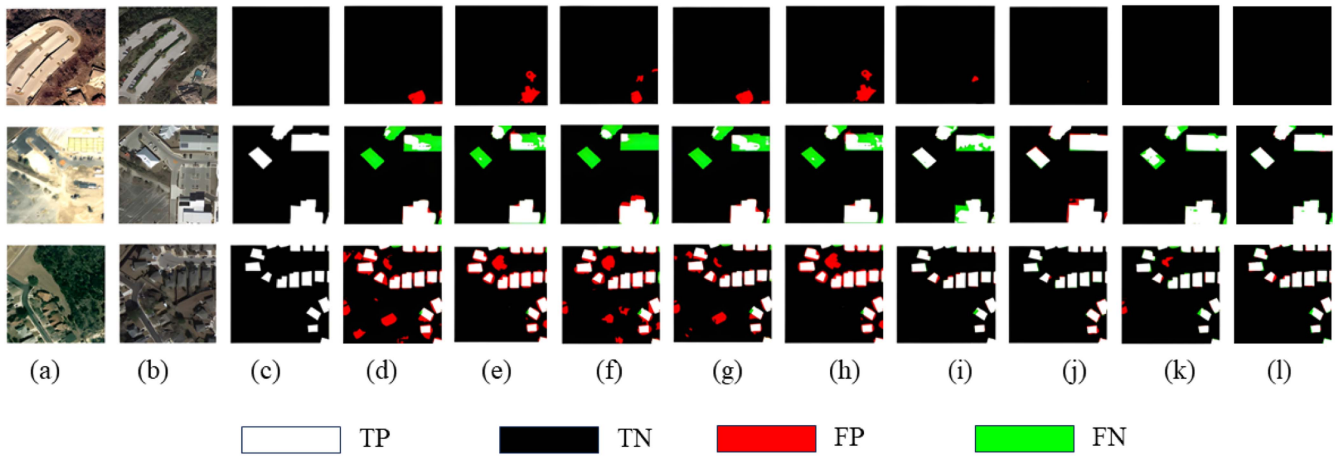

Fig. 6. Visualization of experimental results on the LEVIR-CD. (a) T1 image, (b) T2 image, (c) GT, (d) FC-EF, (e) FC-Siam-Di, (f) FC-Siam-Conc, (g) DTCDSCN, (h) STA, (i) LightCDNet, (j) SGSLN, (k) SEIFNet, (l) CPHSM-Net.

TABLE I
COMPARISON RESULTS ON LEVIR-CD

Model	Pre	Rec	F1	IoU
FC-EF	84.82	77.55	81.02	68.11
FC-Siam-Di	86.73	77.52	81.87	69.31
FC-Siam-Conc	79.85	83.00	81.39	68.62
DTCDSCN	83.12	79.58	81.31	68.51
STA	89.47	83.31	86.28	75.88
LightCDNet	90.60	**91.38**	90.99	83.47
SGSLN	91.08	90.79	90.93	83.38
SEIFNet	91.79	90.21	91.00	83.48
CPHSM-Net	**91.89**	90.28	**91.07**	**83.62**

TABLE II
COMPARISON RESULTS ON WHU-CD

Model	Pre	Rec	F1	IoU
FC-EF	85.97	81.83	83.70	73.72
FC-Siam-Di	87.34	86.33	86.82	77.90
FC-Siam-Conc	85.90	87.68	86.75	77.75
DTCDSCN	91.19	89.21	90.16	82.81
STA	91.11	90.37	90.74	83.05
LightCDNet	90.99	91.73	91.36	83.89
SGSLN	91.51	92.01	91.75	85.15
SEIFNet	**92.20**	90.59	91.39	84.14
CPHSM-Net	91.68	**93.35**	**92.56**	**86.15**

TABLE III
COMPARISON RESULTS ON DSIFN-CD

Model	Pre	Rec	F1	IoU
FC-EF	61.80	57.75	59.71	42.56
FC-Siam-Di	68.44	58.27	62.95	45.93
FC-Siam-Conc	59.08	62.80	60.88	43.76
DTCDSCN	63.75	55.36	59.26	42.11
STA	51.48	36.40	42.65	27.11
LightCDNet	60.10	56.53	58.26	41.10
SGSLN	**76.24**	56.18	64.69	47.81
SEIFNet	64.48	64.98	64.73	47.85
CPHSM-Net	61.82	**70.22**	**65.75**	**48.98**

metrics are described as follows:

$$\text{Pre} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

$$\text{Rec} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

$$\text{F1} = \frac{2 \times \text{Pre} \times \text{Rec}}{\text{Rec} + \text{Pre}} \tag{10}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}}. \tag{11}$$

TP (True Positive) signifies the count of positive class instances that the model accurately identifies as truly existing. Conversely, TN (True Negative) denotes the count of negative class instances that the model accurately determines as absent. FP (False Positive) represents the number of negative class samples that the model falsely identifies as belonging to the positive class. FN (False Negative) refers to the number of positive class instances that the model erroneously detects as belonging to the negative class.

C. Implementation Details

The experiments are all set up based on the Pytorch framework. Training was performed on three Tesla V00 GPUs with the batch size set to 8. The initial learning rate was a set to 0.01 with 0.1 decay every 40 epochs. Also, the parameters are optimized using stochastic gradient descent with momentum of 0.9 and weight decay of 0.0005. Data augmentation is used in the experiments to achieve higher accuracy, including rescaling,

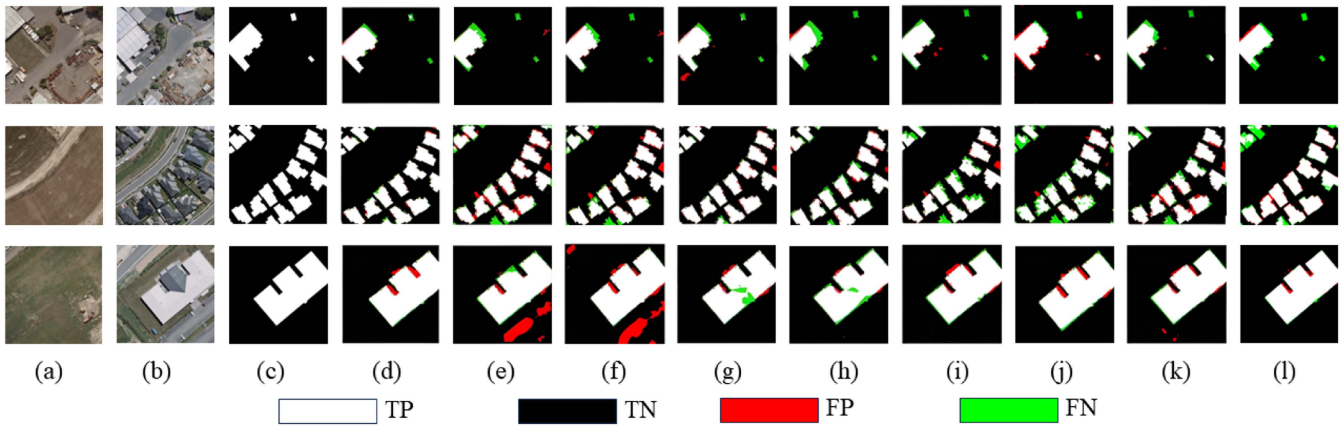

Fig. 7. Visualization of experimental results on the WHU-CD. (a) T1 image, (b) T2 image, (c) GT, (d) FC-EF, (e) FC-Siam-Di, (f) FC-Siam-Conc, (g) DTCDSCN, (h) STA, (i) LightCDNet, (j) SGSLN, (k) SEIFNet, (l) CPHSM-Net.

cropping, flipping, and Gaussian blurring during training, and a validation dataset is used to select the best training weights.

To validate the proficiency of CPHSM-Net in remote sensing image CD, we conduct experiments comparing its performance against other state-of-the-art CD models, namely FC-EF, FC-Siam-Di, FC-Siam [27], STANet [36], DTCDSCN [21], LightCDNet [46], SGSLN [47], SEIFNet [48]. Table I –III show the experimental results on the above three datasets, with the optimal values bolded. The visualization results for the three datasets are depicted in Figs. 6–8. In these figures, white denotes TP, black denotes TN, red denotes FP, and green denotes FN.

D. Comparison and Analysis on LEVIR-CD

Table I indicates that CPHSM-Net performs well in the Pre, F1 coefficient, and IoU metrics, with an improvement of 91.89%, 91.07%, and 83.62%, respectively. However, CPHSM-Net performs less well in the Rec metric, with LightCDNet obtaining the highest Rec, 91.38%. The distance between the front and rear viewpoints of the remote sensing image CD map is considerable. The Pre and Rec indicators are susceptible to a multitude of external factors, whereas the F1 indicator is more objective and therefore more suitable for evaluating the model. Overall, CPHSM-Net outperforms all the comparison algorithms on this dataset.

The result visualization is shown in Fig. 6, where pseudovariable targets, large targets, and dense small targets are randomly selected for visualization. Methods (i)–(l) are effective in coping with vegetation changes, while (k)–(l) are effective in coping with pseudovariations caused by light. This demonstrates that SEIFNet and CPHSM-Net effectively avoid the false detection of pseudochanges.When compared with SGSLN, CPHSM-Net still exhibits some FN performance in the change processing of large targets. However, a comparison of the visualization results indicates that CPHSM-Net is more effective in detecting the edges of targets. For dense small targets, CPHSM-Net demonstrates a superior detection effect and is capable of displaying more regular and complete buildings.

E. Comparison and Analysis on WHU-CD

Table II is derived from the comparison test. Among the models, SEIFNet achieved the highest Pre of 92.20%, with CPHSM-Net ranking second at 91.68%. CPHSM-Net demonstrated the most robust performance across all remaining metrics. F1, which incorporates both Pre and Rec, offers a more comprehensive evaluation of network performance. CPHSM-Net achieves the highest F1 score, in part due to its 93.35% Rec. The WHU-CD dataset is a high-resolution (0.2 m/pixel) dataset. In this dataset, CPHSM-Net is able to efficiently extract and utilise richer information when processing the extracted features, such as CP. Consequently, the leakage detection rate is reduced, and a higher recall value is obtained.

Fig. 7 illustrates the comparative experimental results. Irregular targets, dense small targets, and regular large targets are taken as visualization objects, respectively. It can be observed that CPHSM-Net is capable of effectively dealing with various pseudovariations and obtaining clear and accurate target edges. When recognizing dense small targets, the comparison algorithms exhibit varying degrees of ambiguity and stickiness, which CPHSM-Net effectively addresses. Similarly, when detecting regular large targets, CPHSM-Net is capable of more accurately capturing the target contour.

F. Comparison and Analysis on DSIFN-CD

Table III demonstrates the efficacy of CPHSM-Net, exhibiting superior performance across all metrics compared to other methods. SGSLN exhibits the highest Pre at 76.24%, while CPHSM-Net exhibits the highest Rec at 70.22%. A plausible rationale is that CPHSM-Net places a greater emphasis on the significance of spatial location by integrating spatial and channel information through the ASCE strategy in the feature extraction stage, and subsequently utilizing the spatial information to perform similarity metrics on the features. Although CPHSM-Net performs slightly better than SEIFNet in terms of metrics, SEIFNet achieves a better balance between Pre and Rec. Combining Tables I and II, it can be seen that the accuracies of the individual methods become significantly lower on the DSIFN dataset. The

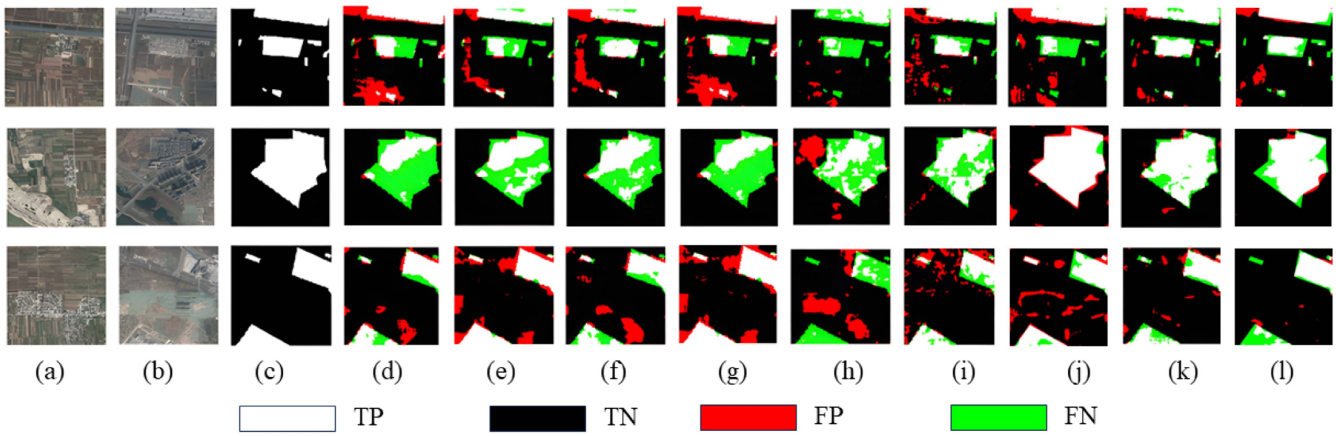

Fig. 8. Visualization of experimental results on the DSIFN-CD. (a) T1 image, (b) T2 image, (c) GT, (d) FC-EF, (e) FC-Siam-Di, (f) FC-Siam-Conc, (g) DTCDSCN, (h) STA, (i) LightCDNet, (j) SGSLN, (k) SEIFNet, (l) CPHSM-Net.

accuracies of the three datasets, DSIFN-CD (2 m/pixel), LEVIR-CD (0.5 m/pixel), and WHU-CD (0.2 m/pixel), are gradually improved. Obviously, the higher the accuracy of the dataset, the network can process the image more clearly. And the dataset is labeled as a large-scale area division rather than a fine-grained division for each house and building. CPHSM-Net focuses on the spatial and channel information of the features and has to interact with the dual time-phase features. Therefore, PHSM-Net can handle high resolution and high precision remote sensing images more effectively. Furthermore, this demonstrates that as remote sensing image technology continues to evolve, our network is poised to gain a significant advantage.

The results are presented in Fig. 8, which shows the visualization of three categories of targets: broken small targets, irregular large targets, and more regular targets. The reduction of pseudochange interference has been a major objective of the CD task. Methods (d)–(h) are unable to effectively deal with "pseudochanges" such as vegetation changes when they exist. Methods (i)–(l) ameliorate this problem, with CPHSM-Net being the most effective. The complete detection of large targets and the accurate localisation of small targets are critical issues when the building scale changes. CPHSM-Net is able to detect large targets more completely, but it requires improvement in edge processing and the detection of small targets.

TABLE IV
ABLATION EXPERIMENTS RESULTS ON THE LEVIR-CD DATASET

Model	Pre	Rec	F1	IoU
Net-1	88.52	91.79	90.12	82.03
Net-2	90.78	90.59	90.68	82.96
Net-3	85.68	**93.23**	89.30	80.66
CPHSM-Net	**91.89**	90.28	**91.07**	**83.62**

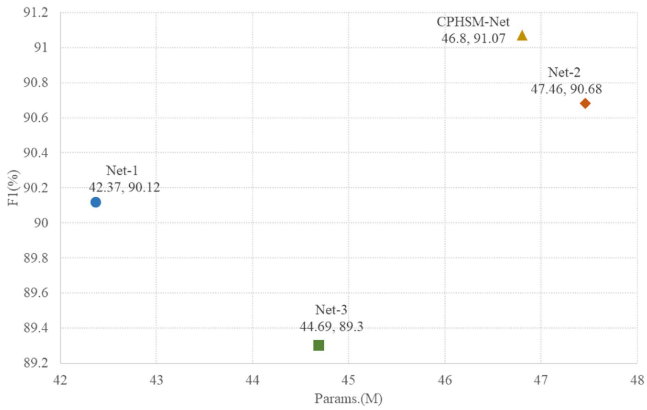

Fig. 9. Findings from ablative trials on the LEVIR-CD dataset.

G. Ablation Study

In order to assess the effectiveness of each component in the CPHSM-Net model, this study conducted ablation experiments on the LEVIR-CD dataset, systematically varying the modules to evaluate their respective performance. Net-1 contains only ASCE; Net-2 contains only CP; and Net-3 contains only HSM. All other settings are consistent with the full CPHSM-Net. Table IV shows the experimental results.Among the networks, Net-3, which contains only the HSM module, obtained the highest Rec. The HSM module divides the features into high and low level features, which are fused separately. Furthermore, it utilizes spatial information through similarity metrics to reduce missed detection, thus enhancing Rec. In comparison to Net-3, Net-1 employs an ASCE strategy in the feature extraction stage, whereby spatial and channel information are utilized in a synergistic manner to enrich the extracted features. Net-1 achieves a certain degree of balance between Pre and Rec, and enhances F1. Net-2 employs a CP module and associations to interactively sense the differences between features, applying attention weights. This process is relatively time-consuming, yet it also leads to an improvement in F1. CPHSM-Net integrates the modules that process the features at different stages, thereby enhancing the overall performance.

The impact of individual modules on model efficiency and model performance is further demonstrated in Fig. 9. The Net-1, which incorporates solely the ASCE strategy, exhibits the lowest

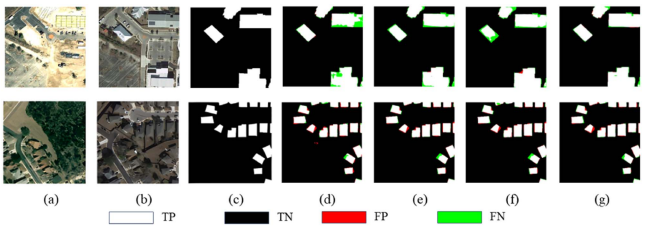

Fig. 10. Visualization of ablation experiments results on the LEVIR-CD dataset. (a) T1 image, (b) T2 image, (c) GT, (d) Net-1, (e) Net-2, (f) Net-3, (g) CPHSM-Net.

number of model parameters. This is due to the fact that the ASCE strategy enriches features without the addition of learning parameters. Net-2, which employs solely the CP module, markedly enhances the model's performance, yet exhibits the lowest efficiency. The CPHSM-Net model, which employs all three methods simultaneously, exhibits an increase in the number of model parameters, yet it demonstrates superior performance. CPHSM-Net employs a combination of modules to achieve a balance between efficiency and performance.

The results of the visualization are presented in Fig. 10. When each method is employed in isolation, an increased number of FP and FN is observed to varying degrees. In the feature extraction phase, it is possible to extract change edge information in an efficient manner by utilizing spatial and channel information in conjunction with the ASCE strategy. However, if this strategy is employed solely within features, the potential for smaller target sticking may also arise. The CP module is capable of interacting with dual time-phase features at varying scales, thereby facilitating the determination of change targets in a more comprehensive manner. The HSM is able to identify and localise minor target boundaries with greater precision, thereby reducing the occurrence of sticking through the utilisation of residual fusion and similarity metrics. As illustrated in Fig. 10, CPHSM-Net effectively integrates these three modules to generate a change map that is most closely aligned with the ground truth map.

H. Comparison and Analysis of Loss Functions

The design and selection of the loss function directly affects the efficiency and effectiveness of model training. Correct selection and design of the loss function can maximize the performance of the model and ensure that it can achieve the expected results in practical applications. In order to further analyze and optimize the model. We select different loss functions for experimentation and analysis. The loss functions employed in this study include the cross-entropy loss function (abbreviated as L-Cross), the focus loss function [49] (abbreviated as L-Focus) and the dice coefficient loss function [50] (abbreviated as L-Dice). The results of the experiment are presented in Fig. 11. In the LEVIR-CD and WHU-CD datasets, the highest F1 score is obtained by using L-Cross training, which outperforms L-Focus and L-Dice. This loss function optimizes the classification accuracy of the model output by directly comparing it with the real category labels through softmax processing of the model output. In the DSIFN-CD dataset, the top-ranked model is L-Focus, with

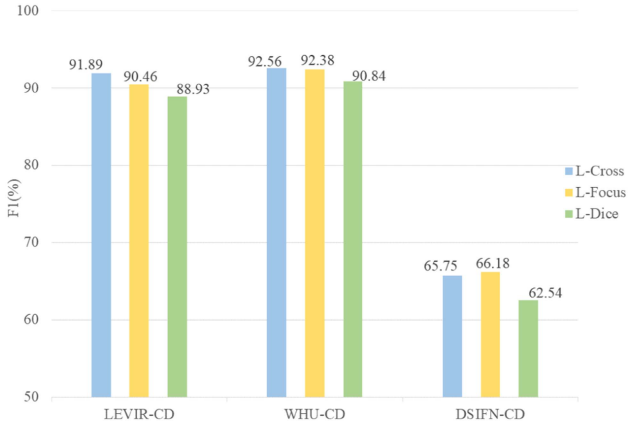

Fig. 11. F1 scores with different loss functions on different datasets.

an F1 score of 66.18%. This dataset contains a variety of surface variations. The Focus loss function mitigates the imbalance between positive and negative sample sizes and classification difficulties. On balance, L-Cross was adopted as the loss function for CPHSM-Net.

V. Discussion

The experiment demonstrates that the proposed method CPHSM-Net outperforms the comparison methods. In the comparison experiments, CPHSM-Net exhibits superior detection results for diverse types of building targets. The necessity and effectiveness of each module are also validated by ablation experiments. Standard CNNs are limited in their capacity to extract information, often overlooking significant local specifics, a pivotal aspect for the CD task. The ASCE strategy effectively leverages spatial and channel information to enrich the extracted features. The CP module, inspired by Transformer, employs an attention mechanism to focus on important information without increasing the number of learnable parameters. In contrast to self-attentive structures, it cross-perceives the bitemporal features and outputs features containing interaction information. The HSM module fuses the processed bitemporal features by residual fusion and similarity metrics. Further utilizing spatial information to improve the extraction of details is an effective approach that is easy to understand.

Nevertheless, there is still considerable scope for enhancement in our methodology.

1) CPHSM-Net, inspired by the transformer, reduces the redundancy of information when processing features. Nevertheless, this results in the omission of some useful data in the DSIFN-CD dataset, which generates GT maps according to the regionally divided change section. This phenomenon is also evidenced by the experimental results, which indicate that the accuracy of the CPHSM-Net on the DSIFN-CD dataset is slightly lower than that of other models. To address this challenge and enhance the generalizability of the method, further improvements are particularly important. In future research, the feature processing strategies for CPHSM-Net must be optimized in order to more effectively retain and utilize key information from

different datasets. This entails the modification and enhancement of the model architecture and the loss function, with a view to achieving better balance between the preservation of information and the reduction of redundancy. It is anticipated that these enhancements will facilitate the utilization of CPHSM-Net in a broader range of practical applications, thereby enhancing its overall applicability and competitiveness.

2) The current model employs a cross-entropy loss function. Despite its simplicity and efficacy, this loss function also exhibits certain shortcomings, including the potential for sensitivity localization and easy localization. The loss function in the CD task primarily addresses two significant challenges, namely sample imbalance and target edge blurring. Many contemporary methods have devised effective loss functions to enhance the performance of their respective methods. In this study, we draw inspiration from methods such as SEIFNet and employ a combination of diverse types of loss functions. The utilization of disparate loss functions is advantageous insofar as they exhibit distinct characteristics. The subsequent objective is to examine the optimal utilisation and configuration of the loss functions, with the aim of further enhancing the CPHSM-Net performance.
3) The CPHSM-Net model structure is rendered more intricate by virtue of its multilayer feature extraction and interrelationship extraction capabilities. This, in turn, gives rise to an increased demand for computational resources and a lengthening of the training time. This complexity is in part attributable to the considerable number of parameters and intricate computational graphs that it is required to process. This represents a significant limitation of CPHSM-Net. It is therefore evident that a reduction in the number of parameters and an increase in computational efficiency represent a crucial avenue for optimising CPHSM-Net. In forthcoming work, the design of lightweight models, such as compressed models, may be considered as a means of reducing the number of parameters, while maintaining the model performance. Furthermore, optimizing the computational flow and algorithmic implementation of the model to enhance the efficiency of training and inference represents a crucial avenue for improvement. It is anticipated that enhancements will diminish the computational complexity of CPHSM-Net, thereby enhancing its applicability in resource-constrained environments and accelerating its training and inference speeds on large-scale datasets.

VI. Conclusion

This article presents a CPHSM-Net for remote sensing image CD. CPHSM-Net fully utilizes spatial and channel information and interacts with the features of dual time-phase imagery. A comparison is made between CPHSM-Net and other state-of-the-art algorithms on the LEVIR-CD, WHU-CD, and DSIFN-CD datasets. The performance of CPHSM-Net is benchmarked against leading algorithms, highlighting its superiority. This model offers a viable solution to address the critical limitations of current CD techniques. The ASCE strategy synergistically integrates spatial and channel information to enhance feature information. The CP module integrates global information through CP between features. The HSM module fuses features into high and low level features, effectively utilizing spatial information through similarity metrics to capture change differences. Experimentally, CPHSM-Net has been demonstrated to be effective in CD tasks, thereby illustrating that it can serves as an instrument for remote sensing applications, such as the detection of urban expansion. It is also important to note that there is still considerable scope for improvement in the proposed method. As a result, our future endeavors will primarily concentrate on refining the model architecture, bolstering its generalization capabilities, and elevating its computational proficiency.

References

[1] M. Noman, M. Fiaz, H. Cholakkal, S. Khan, and F. S. Khan, "ELGC-Net: Efficient local–global context aggregation for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4701611.
[2] M. Kucharczyk and C. H. Hugenholtz, "Remote sensing of natural hazard-related disasters with small drones: Global trends, biases, and research opportunities," *Remote Sens. Environ.*, vol. 264, 2021, Art. no. 112577.
[3] H. Luo, C. Liu, and C. Wu, "Urban change detection based on Dempster–Shafer theory for multitemporal very high-resolution imagery," *Remote Sens.*, vol. 10, no. 7, 2018, Art. no. 980.
[4] J. Yin et al., "Integrating remote sensing and geospatial Big Data for urban land use mapping: A review," *Int. J. Appl. Earth Observation Geoinformation*, vol. 103, 2021, Art. no. 102514.
[5] A. H. Chughtai, H. Abbasi, and I. R. Karas, "A review on change detection method and accuracy assessment for land use land cover," *Remote Sens. Appl.: Soc. Environ.*, vol. 22, 2021, Art. no. 100482.
[6] Z. Y. Lv, W. Shi, X. Zhang, and J. A. Benediktsson, "Landslide inventory mapping from bitemporal high-resolution remote sensing images using change detection and multiscale segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1520–1532, May 2018.
[7] N. Hassiba and Y. Chibani, "Multiple support vector machines for land cover change detection: An application for mapping urban extensions," *ISPRS J. Photogrammetry Remote Sens.*, vol. 61, pp. 125–133, 2006.
[8] M. Hussan, D. Chen, and A. Cheng, "Change detection from remotely sensed images: From pixel-based to object-based approaches," *SPRS J. Photogrammetry Remote Sens.*, vol. 80, pp. 91–106, 2013.
[9] Z. Li, W. Shi, H. Zhang, and M. Hao, "Change detection based on Gaborwavelet features for very high resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 783–787, May 2017.
[10] J. S. Deng et al., "PCA-based land-use change detection and analysis using multitemporal and multisensory satellite data," *Int. J. Remote Sens.*, vol. 29, pp. 23–38, 2008.
[11] L. Yann et al., "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, Dec. 1989.
[12] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6000–6010.
[13] L. Zhang and X. Zheng, "Random forest-based change detection for remote sensing images using multi-feature differences," *Cartogr. Geographic Inf. Sci.*, vol. 34, pp. 149–152, 2022.
[14] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, "Vision transformers for remote sensing image classification," *RemoteSens.*, vol. 13, 2021, Art. no. 516.
[15] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.
[16] Y. Ding et al., "Multi-scale receptive fields: Graph attention neural network for hyperspectral image classification," *Expert Syst. Appl.*, vol. 223, 2023, Art. no. 119858.
[17] J. Feng, Z. Gao, R. Shang, X. Zhang, and L. Jiao, "Multi-complementary generative adversarial networks with contrastive learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5520018.

[18] J. Feng et al., "Class-aligned and class-balancing generative domain adaptation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5509617.
[19] J. Raveerat et al., "Newly built construction detection in SAR images using deep learning," *Remote Sens.*, vol. 11, no. 12, 2019, Art. no. 1444.
[20] H. Jiang et al., "PGA-SiamNet: Pyramid feature-based attention-guided Siamese network for remote sensing orthoimagery building change detection," *Remote Sens.*, vol. 12, 2020, Art. no. 484.
[21] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep Siamese convolutional network model," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 811–815, May 2021.
[22] X. Peng, R. Zhong, Z. Li, and Q. Li, "Optical remote sensing image change detection based on attention mechanism and image difference," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7296–7307, Sep. 2021.
[23] C. Zhang et al., "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 166, pp. 183–200, 2020.
[24] Z. Chen, Y. Song, Y. Ma, G. Li, R. Wang, and H. Hu, "Interaction in transformer for change detection in VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3000612.
[25] W. G. C. Bandara and V. M. Patel, "A transformer-based Siamese network for change detection," in *Proc. IGARSS 2022-2022 IEEE Int. Geosci. Remote Sens. Symp.*, Kuala Lumpur, Malaysia, 2022, pp. 207–210.
[26] W. Shi, M. Zhang, R. Zhang, S. Chen, and Z. Zhan, "Change detection based on artificial intelligence: State-of-the-art and challenges," *Remote Sens.*, vol. 12, no. 10, May 2020, Art. no. 1688.
[27] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional Siamese networks for change detection," in *Proc. IEEE Int. Conf. Image Process.*, 2018, pp. 4063–4067.
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.
[30] Y. Zhao, P. Chen, Z. Chen, Y. Bai, Z. Zhao, and X. Yang, "A triple-stream network with cross-stage feature fusion for high-resolution image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5600417.
[31] Z. Lv, P. Zhong, W. Wang, Z. You, and N. Falco, "Multiscale attention network guided with change gradient image for land cover change detection using remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 2501805.
[32] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5604816.
[33] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[34] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surveys*, vol. 54, pp. 1–41, Jan. 2022.
[35] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. Conf. Workshop Neural Inf. Process. Syst.*, 2014, pp. 1–9.
[36] H. B. Bi, D. Lu, H. H. Zhu, L. N. Yang, and H. P. Guan, "STA-Net: Spatial–temporal attention network for video salient object detection," *Int. J. Speech Technol.*, vol. 51, no. 6, pp. 3450–3459, Jun. 2021.
[37] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[38] K. Han et al., "A survey on visual transformer," 2020, *arXiv:2012.12556*.
[39] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[40] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *2021 IEEE/CVF Int. Conf. Comput. Vis.*, Montreal, QC, Canada, 2021, pp. 9992–10002.
[41] S. Mascarenhas and M. Agarwal, "A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for image classification," in *Proc. 2021 Int. Conf. Disruptive Technol. Multi-Disciplinary Res. Appl.*, Bengaluru, India, 2021, pp. 96–99.
[42] N. S. V. Rao, "Nadaraya-Watson estimator for sensor fusion problems," in *Proc. Int. Conf. Robot. Automat.*, Albuquerque, NM, USA, vol. 3, 1997, pp. 2069–2074.
[43] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proc. 2019 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 3080–3089.
[44] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens.*, vol. 12, no. 10, 2020, Art. no. 1662.
[45] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.
[46] Y. Xing, J. Jiang, J. Xiang, E. Yan, Y. Song, and D. Mo, "LightCDNet: Lightweight change detection network based on VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 2504105.
[47] S. Zhao, X. Zhang, P. Xiao, and G. He, "Exchanging dual encoder-decoder: A new strategy for change detection with semantic guidance and spatial localization," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4508016.
[48] Y. Huang, X. Li, Z. Du, and H. Shen, "Spatiotemporal enhancement and interlevel fusion network for remote sensing images change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5609414.
[49] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.
[50] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Proc. Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*, 2017, pp. 240–248.

Haicheng Qu (Member, IEEE) received the B.S. degree in computer science from Qingdao University of Technology, Qingdao, China, in 2005, the M.S. degree in computer application technology from Liaoning Technical University, Fuxin, China, in 2008, and the Ph.D. degree in information and communication engineering from Harbin Institute of Technology, Harbin, China, in 2016.

He is currently an Associate Professor with the School of Software, Liaoning Technical University. His research interests include remote sensing imagerapid processing and intelligent Big Data processing.

Lijuan Zhang received the B.S. degree in software engineering from Changchun University, Changchun, China, in 2019. She is currently working toward the Master's degree in software engineering from Liaoning Technical University, China.

Her main research interests are computer vision, remote sensing image processing and analysis, especially change detection.