Hyperboloid-Embedded Siamese Network for Change Detection in Remote Sensing Images

Qian Yang, Shujun Zhang, Jinsong Li, Yukang Sun, Qi Han, and Yuanyuan Sun

***Abstract*—Change detection (CD) is a significant branch of remote sensing image analysis. Its main distinction from general semantic segmentation lies in the input of bitemporal images. Recent CD methods have primarily focused on Euclidean space, disregarding the hidden non-Euclidean details due to the high imaging altitude and complex scenes in remote sensing imagery. This limitation hampers the model's performance. In recent years, hyperbolic space has gradually been introduced into deep learning as a classical non-Euclidean space. To explore the potential impact of hyperbolic information in CD, a hyperbolic-embedding-based Siamese network is proposed in this article. Specifically, we propose a hyperbolic similarity attention mechanism that can map deep features from bitemporal images into hyperbolic space, then establish the relationship between bitemporal features based on the hyperbolic distance, fully mining non-Euclidean information, and fusing features from both branches to enhance feature cohesiveness. Furthermore, we design a differential feature enhanced module in the decoder, which utilizes differential operations at multiple scales to highlight the essence of each layer's features and improve feature richness. Experimental results on two public very-high-resolution CD datasets demonstrate that the proposed network achieves better detection accuracy than other state-of-the-art CD methods.**

***Index Terms*—Attention mechanism, change detection (CD), hyperbolic geometry, remote sensing (RS) images.**

I. Introduction

The objective of change detection (CD) in remote sensing (RS) images is to identify semantic changes between a pair of images in the same area at different times. Depending on the training data used, change detection models can be applied in various fields, such as land use [1], environmental monitoring [2], building and forest cover change analysis [3], urban planning [4], and disaster assessment [5]. RS images differ from natural images in several distinct aspects, including high resolution, multispectral, and distinctive viewing angles. Furthermore, with the advancement of computer and sensor technologies, the resolution of RS images has been steadily increasing in recent years [6]. These factors have brought new advantages and challenges to RS image CD. Effectively extracting the rich information contained in very-high-resolution (VHR) RS images and capturing the bitemporal change features required for CD, while ensuring the robustness of the results, has been a critical concern in the relevant field.

In recent years, deep learning methods have experienced rapid development and have been widely applied in the field of RS [7], [8], [9], [10]. In the context of CD tasks, deep belief networks [11], convolutional neural networks (CNN) [12], and other neural network architectures have gradually replaced traditional CD methods [13], [14]. This is attributed to their exceptional capabilities in deep feature extraction, robustness to noise, and the ability to learn automatically from data without the need for extensive manual intervention. Currently, most deep-learning-based backbone networks for CD employ encoder–decoder architectures such as fully convolutional networks (FCNs) [15] and U-Net [16]. These networks demonstrate outstanding performance in image segmentation tasks, including CD, due to their powerful context extraction ability and deep feature extraction effects. However, CD differs from general image segmentation tasks as it involves the input of a pair of dual-time images. To effectively fuse the information from bitemporal images, researchers have introduced the Siamese network [17] and applied it to RS image CD [18], [19], significantly improving detection accuracy. Subsequently, this idea has been widely adopted and has become a baseline network in CD tasks.

However, existing deep learning models still face some challenges. First, current methods cannot effectively integrate bitemporal features, leading to adverse impacts on detection results due to irrelevant changes. The twin-branch structure of the Siamese network, indeed, enhances the cohesiveness of bitemporal inputs to some extent, but the simple parameter sharing and concatenation operations of Siamese networks cannot effectively fuse the features of bitemporal images and can generate redundant information. In order to overcome these shortcomings, some teams [20], [21], [22], [23] employed a combination of Siamese networks and self-attention mechanism (SAM) [24] to integrate contextual and bitemporal information. However, due to the high computational costs of SAM, it is not suitable for dual-branch inputs in CD tasks.

In addition, existing deep-learning-based CD methods are typically defined in Euclidean space, neglecting the information inherent in non-Euclidean spaces. These CD deep learning models are constrained by the Euclidean geometry, which makes it difficult to extract spatial relationships and hierarchical relationships between pixels [25]. Bronstein et al. [26] demonstrated the existence of highly non-Euclidean properties in RS images. The unique characteristics of RS images, such as high imaging

Manuscript received 17 January 2024; revised 25 March 2024; accepted 10 April 2024. Date of publication 17 April 2024; date of current version 1 May 2024. This work was supported by National Natural Science Foundation under Grant 32301702. *(Corresponding author: Shujun Zhang.)*

The authors are with the School of Data Science, Qingdao University of Science and Technology, Qingdao 266061, China (e-mail: zhangsj@qust.edu.cn; ytz@mails.qust.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3390427

altitude, high resolutions, and complex hierarchical structures, make them susceptible to distortions in Euclidean space. Hyperbolic space, as one of the most typical non-Euclidean spaces, can significantly alleviate these effects by representing RS images on the hyperboloid. For example, it is very difficult to extract the occluded features of the buildings in Euclidean representations, whereas, in hyperbolic representations, these occluded portions can be stretched and highlighted. The significance of this is particularly evident in CD, the change region as the detection target is different from the general semantic segmentation target, with huge differences in shape, size, and type, and it is difficult to extract the deeply hidden non-Euclidean features only in Euclidean space.

Ultimately, most of the existing methods focus on using deep semantic information [18] while shallow features containing fine-grained details and edge information are either utilized through skip connections to assist in upsampling deep features [27] or merged after upsampling features to the same size using multiscale strategies [28], [29]. These methods cannot fully highlight the advantageous information of features at different scales, not only the shallow information, but also the hyperbolic information embedded in the deep features found in our study.

To address the earlier problems, we design a hyperboloid-embedded Siamese network (HES-Net) to incorporate the hidden hyperbolic information from Euclidean space into the neural network, leverage the dual-branch advantage of the Siamese network, and combine it with the hyperbolic space characteristics to deeply integrate bitemporal features. The network is based on the Siamese-UNet architecture. We design a hyperbolic similarity attention mechanism (HSAM) to project the deep Euclidean features extracted by the encoder into the hyperbolic space, highlight the detailed features that are flattened in Euclidean space, and construct the bitemporal feature relationship by combining the hyperbolic distance (HD) with the attention mechanism, which fully fuses the bitemporal information and improves the CD accuracy. Additionally, we design a differential feature enhanced module (DFEM), which subtracts the average feature depth from the multiscale features at the decoder end, synthesizes the advantages of multiscale features, strengthens the shallow fine-grained features and the deep hyperbolic features, and increases the feature richness. Moreover, our HSAM as an independent operation does not increase the number of network parameters and has low computational cost.

To sum up, the main contributions of this article include the following.

1) An HSAM is proposed to project feature vectors onto pseudohyperbolic surfaces and determine feature similarity based on the HD. This mechanism effectively integrates bitemporal features at the same scale, incorporating hyperbolic information into the network.
2) A DFEM is put forward to enhance feature richness by emphasizing the differences between multiscale features and highlighting the advantages of each scale's features.
3) A HES-Net is proposed, which combines Euclidean features and hyperbolic features through the utilization of HSAM and DFEM. Extensive experiments demonstrate that our model achieves higher CD accuracy than other state-of-the-art (SOTA) methods.

The rest of this article is organized as follows. The related works are reviewed in Section II. A detailed description of the proposed method is provided in Section III. The experimental results are reported in Section IV. Discussions of key issues are given in Section V. Finally, Section VI concludes this article.

II. Related Work

A. Backbone Network for RS Image CD

Early deep learning CD methods can be categorized into pixel-based and object-based approaches. Pixel-based methods involve extracting deep features from individual pixels or their neighborhoods in bitemporal images [30], [31], which can be inefficient and prone to noise in high-resolution RS images [32]. On the other hand, object-based methods segment the images into objects and compare them to obtain change maps [33], [34], but they heavily rely on accurate object segmentation algorithms and preprocessing operations, making them susceptible to various factors and lacking robustness [35].

The emergence of FCNs [15] has greatly benefited the field of CD. FCNs possess powerful capabilities in extracting contextual information and can handle image inputs of any size for end-to-end training. This makes them well-suited for scene-based CD tasks. The U-Net [16] is based on FCNs and has the advantage of fusing multiscale features efficiently and accurately. These studies make the encoder–decoder structure the mainstream backbone for DL in the field of CD. Addressing the unique characteristics of the bitemporal inputs in CD, Daudt et al. [19] introduced the shared-parameter Siamese network structure, which enhances the utilization of bitemporal images and significantly improves CD accuracy. It has become the baseline networks for CD [36], [37], [38], [39]. For example, Fang et al. [18] combined the Siamese structure with UNet++ [40], introducing a densely connected Siamese network (SNU-NET), which alleviates the loss of localization information in the deep layers of the neural network through compact information transmission between encoder and decoder. Lei et al. [41] introduced the spatial–spectral feature cooperation (SSFC) strategy based on the Siamese structure to enhance feature richness and so on. However, these models have limitations in modeling context relationships during the convolution process and suffer from the inductive bias of limited local receptive fields. To address these issues, attention mechanisms [24] have been widely adopted as embedding modules in visual tasks to recalibrate convolutional feature maps.

B. Attention Mechanism for RS Image CD

In recent years, the effectiveness of attention mechanisms in various computer vision tasks has been demonstrated. Attention modules have been integrated into existing frameworks for CD tasks. For example, in SNU-NET [18], the enhanced channel attention module is employed to aggregate objects at different scales within the U-Net structure. DASNet proposed by Chen et al. [42] combines channel attention and spatial attention [43] to capture more discriminative features. However, these methods either overly emphasize the fusion of multichannel information at the decoder or focus on the importance of individual channels between pixels, neglecting the interrelationships among the features extracted from bitemporal RS images.

To address these limitations, Chen et al. [21] introduced the BIT model, which incorporates a transformer to encode change regions using semantic tokens. This allows for the fusion of deep features with the original image features to generate change maps. Additionally, Chen et al. [29] designed a scale-aware module to calculate the cross-scale attention of Dis, concentrating the features on more important channels. They further extract the changed pixels through cross-self-attention. These methods consider the interrelationships among features from bitemporal images, but they are still constrained by the extraction of original features, making it difficult to incorporate deeper-level detailed features that are not available in Euclidean space. Moreover, these methods rely on SAM that exhibits exponential computational complexity with increasing input size, making them less suitable for VHR and bitemporal inputs from an application perspective.

C. *Applications of Hyperbolic Space in Deep Learning*

Hyperbolic space is a type of Riemannian space with negative curvature, which enables the representation of hierarchical structures that are not easily captured in Euclidean space. For data with prominent hierarchical characteristics, hyperbolic space can restore their inherent hierarchical structure [44]. Most existing deep learning methods are defined in Euclidean space, benefiting from the isomorphism between Euclidean space and vector space, which facilitates the computation of vectors or matrices required by neural networks [45]. However, vector operations in hyperbolic space need to satisfy manifold constraints, and as a result, they require definition in the gyrovector space, where the representation reflects the characteristics of the manifold.

In the early stages, the applications of hyperbolic space in deep learning were primarily focused on NLP, graph neural networks, knowledge graphs, and similar areas. However, more recent research has discovered the potential of modeling computer vision tasks in non-Euclidean spaces [46]. The physical space we inhabit is a space with curvature, and RS images, which serve as a means to represent the real world, are flattened in Euclidean space. This flattening process causes certain non-Euclidean features to be concealed, making them difficult to be learned by Euclidean networks. Other studies have indicated that expert-designed dissimilarity measures often exhibit non-Euclidean behavior in certain applications [47]. Moreover, hyperbolic models have demonstrated the ability to generate high-quality representations, even in low-dimensional embedding spaces. This characteristic makes them particularly advantageous in scenarios with limited memory and storage [48]. Hyperbolic space has two commonly used representations: 1) the Lorentz model and 2) the Poincaré ball model. Zhang et al. [49] have shown that in the Poincaré ball model, the HD can serve as a metric for uncertainty. By measuring uncertainty or smoothing decision boundaries, we can enhance the robustness of neural networks. In this study, the proposed HSAM is based on this theory and combines HD to construct hyperbolic attention.

III. Methods

A. *Overview*

This section provides a detailed description of our HES-Net. The network is based on a combination of Siamese architecture and U-Net. The method framework is illustrated in Fig. 1. First, a Siamese CNN encoder with shared parameters is employed to extract multilevel features from the bitemporal images. The features from both branches are preliminarily fused using concatenation to prevent information loss. Next, the HSAM is applied to the two deepest layers of the network to map the features into hyperbolic space. By introducing non-Euclidean characteristics and computing the HD as a similarity measure between paired hyperbolic features, the original features are deeply fused using high-fidelity hyperbolic attention and refined with global context. Finally, the features of each layer of the encoder are upsampled, combined, and then passed through the DFEM to highlight the advantages of features at different scales, improve the feature richness, and generate the final pixel-level prediction.

B. *Hyperbolic Similarity Attention Mechanism*

Existing CD methods flatten the extracted features from RS images to satisfy the geometric axioms of Euclidean space. However, the hierarchical structure in RS images is even more complex than natural images. Flattening the features leads to the loss of spatial information, such as image distortions caused by variations in building heights or shadows. To address this issue, we propose the HSAM, which introduces hyperbolic properties into Euclidean-based neural networks without adding extra parameters or computational complexity. The detailed structure of HSAM is illustrated in Fig. 2. The operation of HSAM can be divided into two steps: 1) calculating HD and 2) constructing similarity attention.

1) Feature Projection and HD Calculation: To perform operations in hyperbolic space, we first need to project the original Euclidean features extracted by the Siamese network onto the hyperbolic surface $\mathbb{D}_c^d$. In hyperbolic geometry, this mapping is known as the exponential map, as shown in

$$y = \exp_x^c(v) \tag{1}$$

where $x, y \in \mathbb{D}_c^d$, x is the anchor point and $1/c$ is the hyperbolic radius; $v \in \mathbb{R}^n \cong T_x\mathbb{D}_c^d$ is a vector in Euclidean space; $T_x\mathbb{D}_c^d$ is the tangent space of the hyperbolic surface at point x.

According to the chosen hyperbolic model, different mapping functions are required. James et al. [50] describe five common hyperbolic models. In this study, we utilize the Poincaré ball model. The Poincaré ball is a spherical model in hyperbolic space, where the hyperbolic space can be embedded into a higher-dimensional Euclidean space and represented using the Poincaré ball. When we map a Euclidean space vector to the hyperbolic space, we essentially map it to a point on the Poincaré

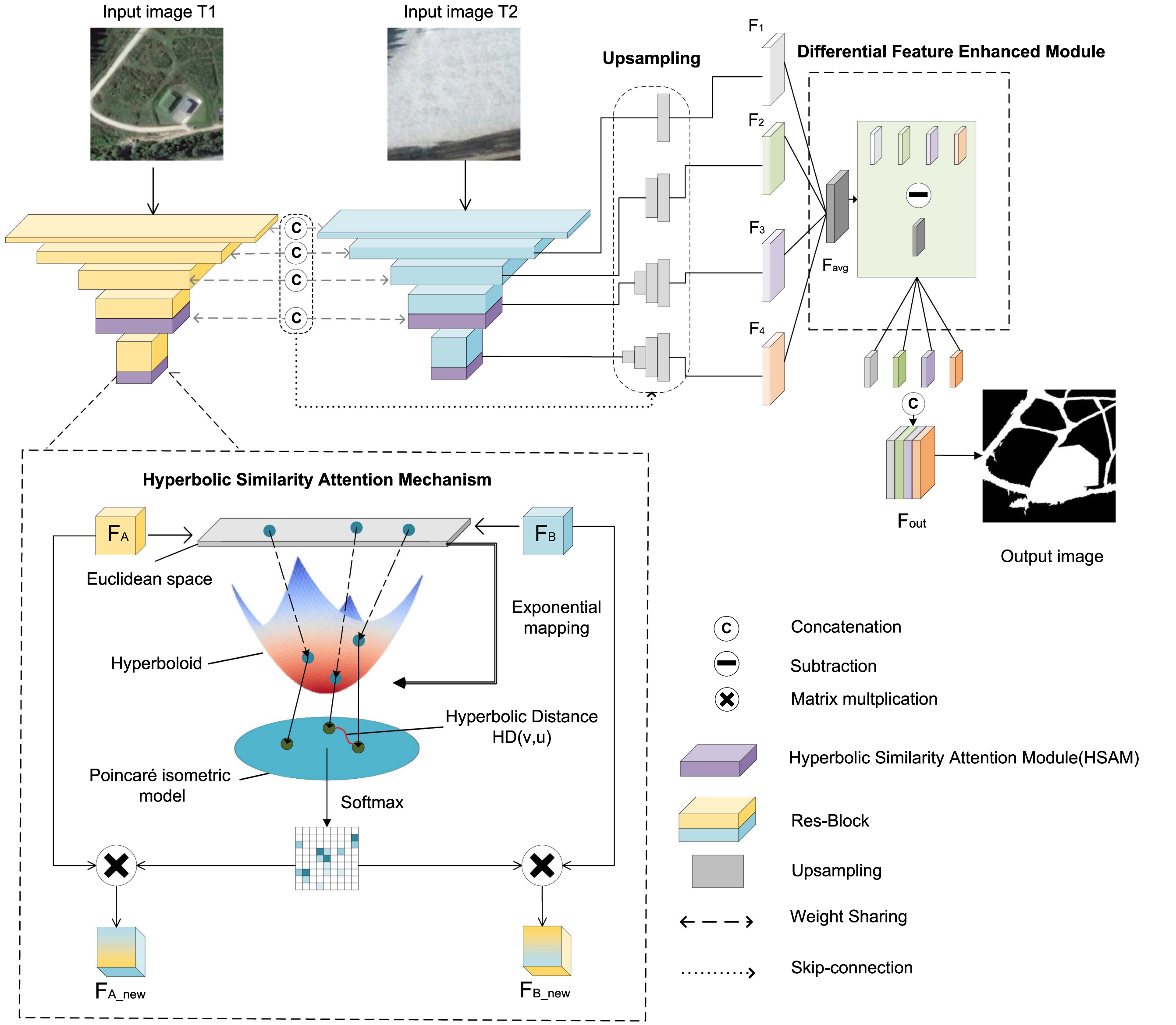

Fig. 1. Overall architecture of the proposed HES-Net. (a) Backbone of HES-Net. The results of each layer after concatenation, according to different scales, are connected to the upsampled results through skip-connections. (b) HSAM integrates deep bitemporal features and embeds hyperbolic information. (c) Differential feature enhanced module (DFEM) highlights shallow fine-grained features and deep hyperbolic features through subtraction operations.

ball. This point in the Poincaré ball model corresponds to a point in the hyperbolic space. The exponential mapping formula in the Poincaré ball model is defined as

$$\exp_x^c(v) = x \oplus_c \left(\tanh\left(\frac{\sqrt{c}\lambda_x^c}{2} \|v\| \right) \frac{v}{\sqrt{c}\|v\|} \right) \tag{2}$$

where $\lambda_x^c = \frac{2}{1-c\|x\|^2}$ is the conformal factor that scales the local distances, $\|\cdot\|$ is the norm l_2, and $\oplus_c$ is the Mobius addition.

In practical applications, we set the anchor point x at the origin and the radius parameter c as 1, resulting in (2) being transformed as follows:

$$\exp_0^1(v) = \frac{\tanh(\|x\|) \cdot x}{\|x\| + eps} \tag{3}$$

where eps is a very small value used to prevent division by zero.

The process of projecting and labeling a pair of Euclidean space feature vectors can be described as follows:

$$F_{jH} = \mathrm{MT}(\exp_0^1(F_j)) \tag{4}$$

where $F_j \in \mathbb{R}^{C\times H\times W}, j \in (1,2)$ represents a pair of original feature vectors at the same scale. $F_{jH} \in \mathbb{R}^{C\times H\times W}$ is the gyrovector obtained by projecting F_j onto the hyperbolic space. $\mathrm{MT}(\cdot)$ represents the operation of constructing a manifold tensor, which is used to sequentially store the gyrovectors.

Afterward, the obtained pair of hyperbolic gyrovectors, F_{1H}, F_{2H}, are reshaped to $\mathbb{R}^{HW\times C_h}$ and $\mathbb{R}^{C_h\times HW}$, respectively, in order to calculate the HD between them. C_h represents the dimensionality of the gyrovector. The HD can serve as a metric for measuring the similarity between the two gyrovectors. When used in CD tasks, the HD captures the underlying

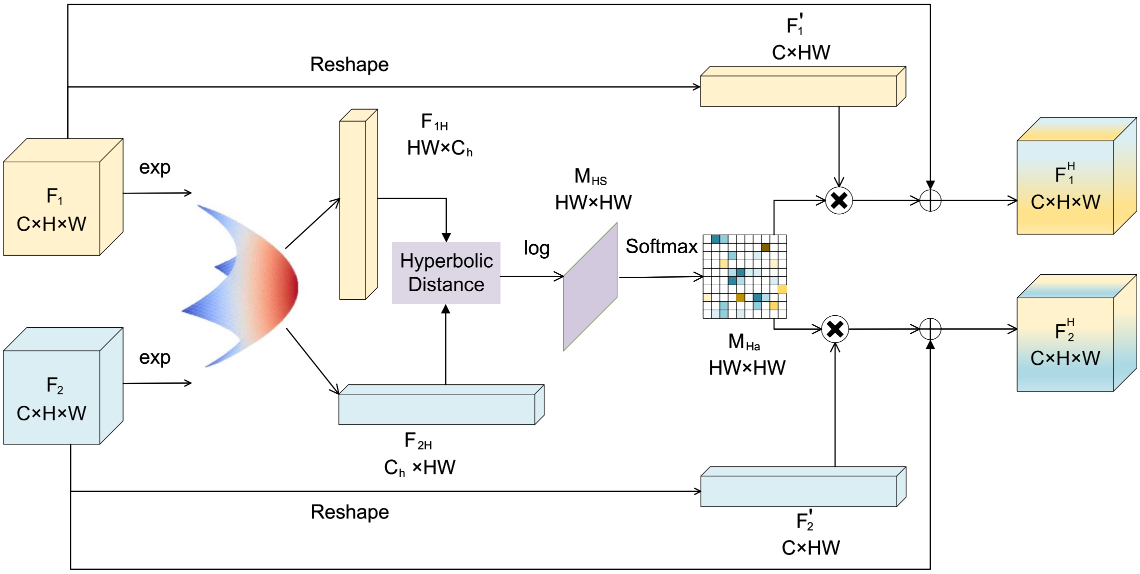

Fig. 2. Illustration of HSAM, where "exp" and "log" represent exponential and logarithmic mapping operations, "⊕" implies the elementwise summation, and "⊗" implies the multiplication operation.

hyperbolic properties within RS images and represents the similarity between the bitemporal inputs at the corresponding spatial locations. By using the HD as a metric and constructing attention maps, it allows for compensating the missing hyperbolic information in Euclidean features. This facilitates the deep fusion of bitemporal input features and emphasizes the regions of change. The calculation of the HD can be represented as follows:

$$\mathrm{HD}(v,u) = \cosh^{-1}\left(1 + \frac{2\left\|v-u\right\|^2}{(1-\left\|v\right\|^2)(1-\left\|u\right\|^2)}\right) \tag{5}$$

where v and u are the hyperbolic gyrovectors corresponding to the same spatial location in the pair of bitemporal data. They are defined on the same Poincaré ball. Therefore, the HD can be normalized to indicate the hyperbolic similarity between the bitemporal features at a specific location.

2) Constructing Hyperbolic Similarity Attention: To obtain the hyperbolic similarity attention map, the first step is to normalize the HDs of the gyro vector features F_{1H}, F_{2H} to obtain the hyperbolic similarity matrix M_{HS}

$$M_{\mathrm{HS}} = \frac{\mathrm{HD}(F_{1H}, F_{2H})}{\sqrt{C^h}} \tag{6}$$

where C_h is the dimension of the gyrovector.

Subsequently, the hyperbolic similarity attention map can be obtained by feeding M_{HS} into the SoftMax layer converted to a probability distribution map M_{Ha}. However, since SoftMax is defined in Euclidean space and involves exponentiation and normalization operations, it cannot be directly applied in the hyperbolic space. Therefore, it is necessary to first map the tangent vectors on the manifold to points on the manifold using the exponential mapping, satisfying the flattened Euclidean space characteristics. Then, the SoftMax operation is performed. Finally, the logarithmic mapping is applied to restore the tangent vectors, resulting in the desired attention map. The equation defining the generation of the hyperbolic similarity attention map is as follows:

$$M_{Ha} = \mathrm{logmap}(\mathrm{Softmax}(\mathrm{expmap}(M_{\mathrm{HS}}))) \tag{7}$$

where $\mathrm{expmap}(\cdot)$ represents the exponential mapping and $\mathrm{logmap}(\cdot)$ represents the logarithmic mapping. $M_{Ha} \in \mathbb{R}^{HW\times HW}$ is the attention map that represents the hyperbolic similarity.

The pair of original features, $F_j \in \mathbb{R}^{C\times H\times W}$, where $j \in (1,2)$, are individually split into two branches. One branch from each feature is reshaped and multiplied elementwise with the attention map M_{Ha}, yielding the attention features. The remaining branches of each feature act as residuals and are added to their respective attention features. The final result is the generation of new features that possess hyperbolic properties and integrate the bitemporal information. The equation defining the generation of the new features is as follows:

$$F_j^H = (\mathrm{reshape}(F_j) \times M_{Ha}) \oplus F_j \tag{8}$$

where $\oplus$ is elementwise addition.

Finally, the new features $F_j^H \in \mathbb{R}^{C\times H\times W}$, where $j \in (1,2)$, will replace the original features F_j in the network computations. Experimental results have shown that it is not necessary to apply HSAM to every feature extracted at each layer of the encoder for optimal performance. This is due to the characteristics of the Siamese-UNet network architecture, such as dense skip connections and weight sharing. The hyperbolic properties introduced by F_j^H through HSAM can propagate to all scales as the encoder performs further convolutions and the decoder progressively upsamples the features. In practice, we only applied HSAM to the deepest two layers of the four spatial scales. The rationale behind this choice will be discussed in Section V.

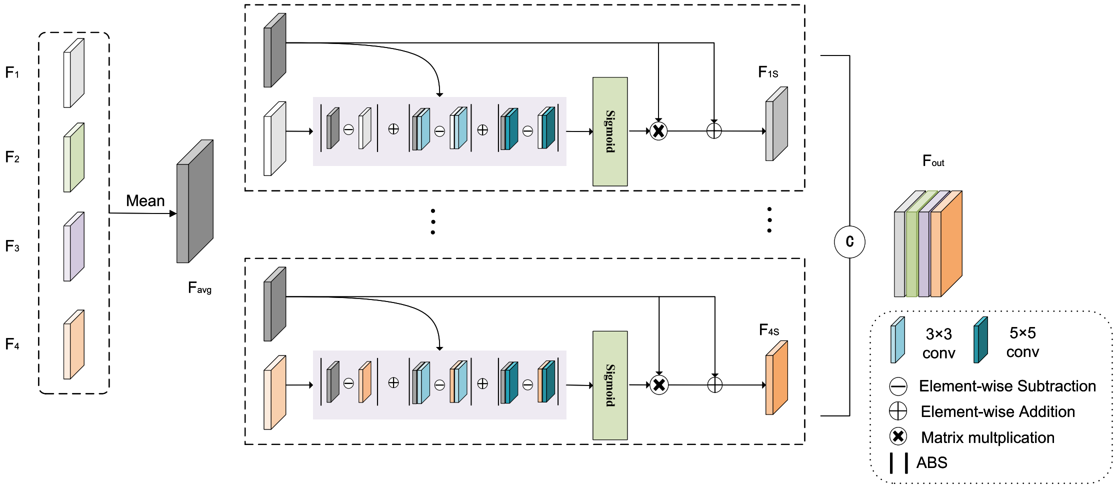

Fig. 3. Structure of the DFEM. Feature averaging ensures keeping the key features while deep subtraction operations can highlight detailed information at different scales.

C. Differential Feature Enhanced Module

Different scales of features often represent different semantic levels due to variations in information density. HES-Net produces four outputs in the decoder, each with the same size as the input image but representing different semantic levels. The shallow outputs contain richer fine-grained features and more precise edge information, but they are also more susceptible to noise. The features from the deeper layers possess richer semantic information and coarse-grained features, allowing for more accurate identification of areas with changes. However, the details of these changes may appear blurred. Additionally, since HSAM is only applied to the deeper layers of the decoder, despite the information propagation facilitated by dense skip connections, the deep features still exhibit more pronounced hyperbolic properties compared to the shallow features. To effectively leverage the advantages of features at different scales, the DFEM is introduced in this study prior to the final output prediction.

The architecture of DFEM is illustrated in Fig. 3. In DFEM, we apply subtraction and feature enhancement operations to the four decoder outputs $F_i, i \in (1, 4)$, in conjunction with the average feature F_{avg}. The subtraction operation between F_i and F_{avg} serves to achieve the effect of image smoothing, helping to mitigate the impact of noise in the shallow features to some extent. Furthermore, F_{avg} highlights the information of common interest across all levels, specifically facilitating the precise localization of changed regions. The formula for obtaining F_{avg} is expressed as

$$F_{\text{avg}} = \frac{\sum_{n}^{i=1} F_i}{n} \tag{9}$$

where n represents the number of decoder output features, in this study, n is set to 4. The notation $\sum \cdot$ denotes elementwise summation.

To enhance the advantages of features at different scales without suppressing edge information, we perform deep subtractions between F_i and F_{avg}. The size of change regions in RS images can vary significantly. To better extract features from different receptive fields, we employ multiscale convolutions of 3×3 and 5×5. Compared to a single-scale differencing module, the multiscale differencing module can capture more information, thereby improving the accuracy of the model. Additionally, the multiscale subnetwork can extract higher-order complementary information, further enhancing the precision of the model. The deep subtraction operation is formulated as

$$\begin{aligned} F_{Di} = \text{Conv}_{3\times3}(&|F_i \ominus F_{\text{avg}}| \oplus \\ &|\text{Conv}_{3\times3}(F_i) \ominus \text{Conv}_{3\times3}(F_{\text{avg}})| \\ &\oplus |\text{Conv}_{5\times5}(F_i) \ominus \text{Conv}_{5\times5}(F_{\text{avg}})|) \end{aligned} \tag{10}$$

where F_{Di} represents the subtraction result of the ith layer, $\ominus$ denotes elementwise subtraction, $\oplus$ represents elementwise addition, and $\text{Conv}_{n\times n}$ denotes a convolutional filter of size $n \times n$. The purpose of adding the three subtraction results is to delve deeper into the exploration of differential information.

Deep subtraction can highlight the advantages of features at different scales, providing richer feature representations. However, using only subtraction may lead to the loss of crucial information. Therefore, after applying subtraction to obtain F_{Di}, we further enhance the features by passing them through a sigmoid activation function and multiplying them with F_{avg}. This helps to enrich the features. To recover the lost information, we employ residual connections. Finally, the enhanced features from each layer are concatenated to generate the final output F_{out}. The formula for obtaining F_{out} is as follows:

$$F_{Si} = F_{\text{avg}} \oplus (\text{Sigmoid}(F_{Di}) \times F_{\text{avg}}) \tag{11}$$

$$F_{\text{out}} = \text{Concat}(F_{1S}, F_{2S}, F_{3S}, F_{4S}) \tag{12}$$

where F_{Si} represents the enhanced features of the ith layer, and Concat denotes the concatenation operation along the channel dimension.

Previous CD methods have typically used subtraction operations only at the encoder side to obtain the difference information between bitemporal inputs. In the decoder, a simple concatenation operation is commonly used to fuse multiscale features. However, this simple concatenation can result in redundant information. In contrast, our proposed DFEM leverages the advantages of multiscale features while reducing information redundancy through subtraction operations. This enables us to obtain more precise localization and richer details in the features. In Sections IV and V, we conducted ablation experiments and effectiveness analysis of DFEM to validate its performance.

D. Loss Function

The CD task in RS images is a subclass of image segmentation tasks, where the final output is a binary image indicating the presence or absence of changes. Because the changed regions are often much smaller in comparison to unchanged regions, CD datasets commonly suffer from class imbalance. To address this issue, we employ a hybrid loss function that combines the widely used dice loss for segmentation tasks and the weighted cross-entropy loss specifically designed to handle class imbalance. Our loss function is defined as follows:

$$L = L_{\text{dice}} + L_{\text{wce}} \tag{13}$$

dice loss and weighted cross-entropy loss are, respectively, defined as

$$L_{\text{dice}} = 1 - \frac{2 \cdot Y \cdot \text{Softmax}(\hat{Y})}{Y + \text{Softmax}(\hat{Y})} \tag{14}$$

where Y indicates the ground truth, $\hat{Y}$ is change map

$$L_{\text{wce}} = \frac{1}{H \times W} \sum_{k=1}^{H \times W} \text{weight}\,[\text{class}] \cdot \left(\log \left(\frac{\exp(\hat{y}\,[k]\,[\text{class}])}{\sum_{l=0}^{1} \exp(\hat{y}\,[k]\,[l])} \right) \right) \tag{15}$$

where $\hat{y}$ is the point in $\hat{Y}$, H, W is the height and width of $\hat{y}$, the value of “class” is 0 or 1, indicating unchanged and changed pixels, respectively.

IV. Experiments

We conduct extensive comparative experiments to evaluate the performance of HES-Net. Furthermore, we perform ablation experiments to validate the effectiveness of HSAM and DFEM.

A. Experimental Setup

1) Datasets: We conducted experiments and analysis using the most authoritative evaluation datasets in the field of CD, namely CDD [51] and LEVIR-CD [52].

The CDD dataset consists of 11 pairs of multispectral images captured by Digital Globe at different seasons, and the spatial resolution of the obtained images ranges from 0.03 to 1 m. A total of 16 000 image pairs each with a size of 256 × 256 are obtained through random cropping and data enhancement, The final dataset was divided into three parts: 1) 10 000 pairs for training, 2) 3000 pairs for validation, and 3) 3000 pairs for testing.

LEVIR-CD dataset consists of 637 pairs of high-resolution RS images from various locations in Texas, USA, with a resolution of 0.5 m and a size of 1024 × 1024 pixels. The dataset is specifically designed for building CD and was captured at different times between 2002 and 2018. To reduce computational complexity and align with recent CD literature, we adopted a patch-based approach. They are cropped into 256 × 256 image pairs by random cropping. This yielded a total of 10 192 images, dividing into 7120 images for training, 1024 images for validation, and 2048 images for testing.

2) Implementation Details: HES-Net was implemented using the PyTorch framework on an NVIDIA RTX A5000 GPU. The model was trained for 120 epochs to achieve full convergence. To highlight the effectiveness of our HSAM and DFEM, and minimize the influence of other factors, we did not employ any data augmentation techniques. The batch size for input images during training was set to 16. We employed the Adam optimizer with a learning rate of 7e-4, and the learning rate was decayed by a factor of 0.8 every 8 epochs. The weights of each convolutional layer were initialized using KaiMing normalization.

3) Evaluation Metrics: In the experiments, we evaluated the model performance using the three commonly used metrics in CD: 1) precision (Pre), 2) recall (Rec), and 3) F1 score (F1). The calculation methods for these metrics are as follows:

$$\text{Pre} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{16}$$

$$\text{Rec} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{17}$$

$$\text{F1} = \frac{2 \times \text{Pre} \times \text{Rec}}{\text{Pre} + \text{Rec}} \tag{18}$$

where false negative (FN) represents the number of pixels that are incorrectly classified as unchanged. False positive (FP) represents the number of pixels that are incorrectly classified as changed. True positive (TP) is the number of pixels correctly classified as changed. The F1, which comprehensively reflects the levels of precision and recall, serves as the primary criterion for evaluating the quality of our model.

B. Comparison Experiments

1) Comparison Methods: We compare our experimental results with the following eight deep-learning-based CD methods, including both classic models in the CD field and the latest models in recent years. These methods, such as our Baseline, mostly adopt a Siamese architecture as follows.

FC-Siam-Conc [19] combines Siamese networks with U-Net by fusing same-scale features before decoding.

STANet [52] incorporates spatiotemporal attention modules to model the spatiotemporal relationship between bitemporal RS images, enhancing feature discrimination for different-scale objects.

FDCNN [53] generates multiscale and multidepth feature difference maps to improve CD accuracy.

TABLE I
PERFORMANCE COMPARISON ON CDD DATASET

Methods	Pre(%)	Rec(%)	F1(%)
FC-Siam-Conc	72.55	57.98	65.21
STANet	88.98	93.11	91.00
FDCNN	87.90	86.99	87.44
DASNet	91.41	92.45	91.92
SNUNet/32	95.61	94.05	94.83
BIT	92.89	94.02	93.45
HMLNet	95.10	93.80	94.40
USSFC-Net	93.65	96.08	94.74
Ours	**97.67**	**96.77**	**97.22**

TABLE II
PERFORMANCE COMPARISON ON LEVIR-CD DATASET

Methods	Pre(%)	Rec(%)	F1(%)
FC-Siam-Conc	74.96	90.53	82.01
STANet	86.14	89.39	87.73
FDCNN	82.99	88.71	85.76
DASNet	84.15	89.78	86.87
SNUNet/32	89.43	87.72	88.57
BIT	89.24	89.37	89.31
HMLNet	84.42	89.91	87.08
USSFC-Net	88.79	**91.17**	89.96
Ours	**90.74**	90.33	**90.53**

The bold values represent the optimal performance for each indicator.

DASNet [42] introduces a dual attention mechanism to penalize features in pseudochange regions and enhance the attention weights of change region features.

SNUNet [18] combines Siamese networks and UNet++ to strengthen information transmission and employs an integrated channel attention module for multiscale feature fusion and improved detail recognition.

BIT [21] introduces a transformer-based encoder to model contextual information and represent high-level semantic features.

HMLNet [54] proposes a hierarchical metric learning network using ensemble learning for same-scale feature learning, enhanced by dual attention modules for internal consistency within change objects.

USSFC-Net [41] employs multiscale decoupled convolution for extracting multiscale features and incorporates spatial–spectral attention for enhanced feature richness.

To ensure fairness, all methods were implemented with the same data partitioning settings, and whenever possible, the original authors' open-source code for the comparative methods was used.

2) Comparison of the CDD Dataset: The quantitative analysis of the experimental results on the CDD dataset is presented in Table I, with the best values for each metric highlighted in bold. The visual analysis is illustrated in the upper half of Fig. 4. Through quantitative analysis, it is evident that our proposed method surpasses the comparative methods in all three metrics. Compared to SNUNet, which is one of the most commonly cited CD methods in the past two years, our method achieves improvements of 2.06%/2.72%/2.39% in Pre, Rec, and F1, respectively. When compared to the latest CNN-based methods, HMLNet and USSFC-Net, our method demonstrates respective improvements of 2.57%/2.97%/2.82% and 4.02%/0.69%/2.48% in Pre, Rec, and F1.

The CDD dataset is a complex target CD dataset with seasonal variations, which imposes demands on the robustness of the network. For example, in the first row of Fig. 4, the road is highlighted by weather changes and becomes a pseudochange area. The second and fourth rows depict complex changes influenced by seasonal factors, where some road changes are highly ambiguous. However, our HES-Net exhibits almost no false detections or omissions compared to other methods, accurately identifying even the ambiguous details. In the third row, which involves changes in small areas, it can be observed that our method's results are closer to the ground truth compared to other methods, faithfully restoring irregular edges like the ground truth. Through visual analysis, it is evident that our method possesses strong noise resistance and fine-detail detection capabilities, with the hierarchical structure of hyperbolic space greatly aiding in the classification of complex scenes.

3) Comparison on the LEVIR-CD Dataset: The quantitative analysis of the experimental results on the LEVIR-CD dataset is presented in Table II, and the visual analysis is illustrated in the lower half of Fig. 4. From Table II, it can be observed that our method achieves satisfactory results on the LEVIR-CD dataset. Compared to the latest transformer-based method, BIT, our method demonstrates improvements of 1.94%, 0.44%, and 1.19% in Pre, Rec, and F1, respectively. Although the recall rate does not reach the SOTA, our method outperforms the latest CNN-based methods, HMLNet and USSFC-Net, by 3.42% and 0.54% in terms of F1, respectively, which better reflects the overall model performance. Moreover, the Pre metric shows significant improvements of 6.76% and 2.99%, respectively.

The LEVIR-CD dataset focuses on dense building CD. As shown in the second row of the lower half of Fig. 4, our method produces detection results in dense urban areas that are closer to the ground truth compared to other methods, especially in terms of capturing edge shapes. In both the first and fourth rows, which depict large-scale building changes, our method effectively segments different buildings, demonstrating its ability to recognize fine-grained edge details. In the top-left of the fourth row, there is a region with pseudochanges that are prone to false alarms. While most other methods mistakenly detect changes in this area, our method accurately identifies the pseudochange region. This highlights the excellent capability of our HES-Net in capturing details in specific regions. Overall, our model exhibits the best performance on both datasets.

C. Ablation Study

To validate the effectiveness of HSAM and DFEM, we conducted a series of ablation experiments using different combinations of modules on both datasets. The results of these experiments are presented in Tables III and IV.

As shown in Tables III and IV, our ablation experiments exhibit consistent trends on both datasets. Our baseline model is

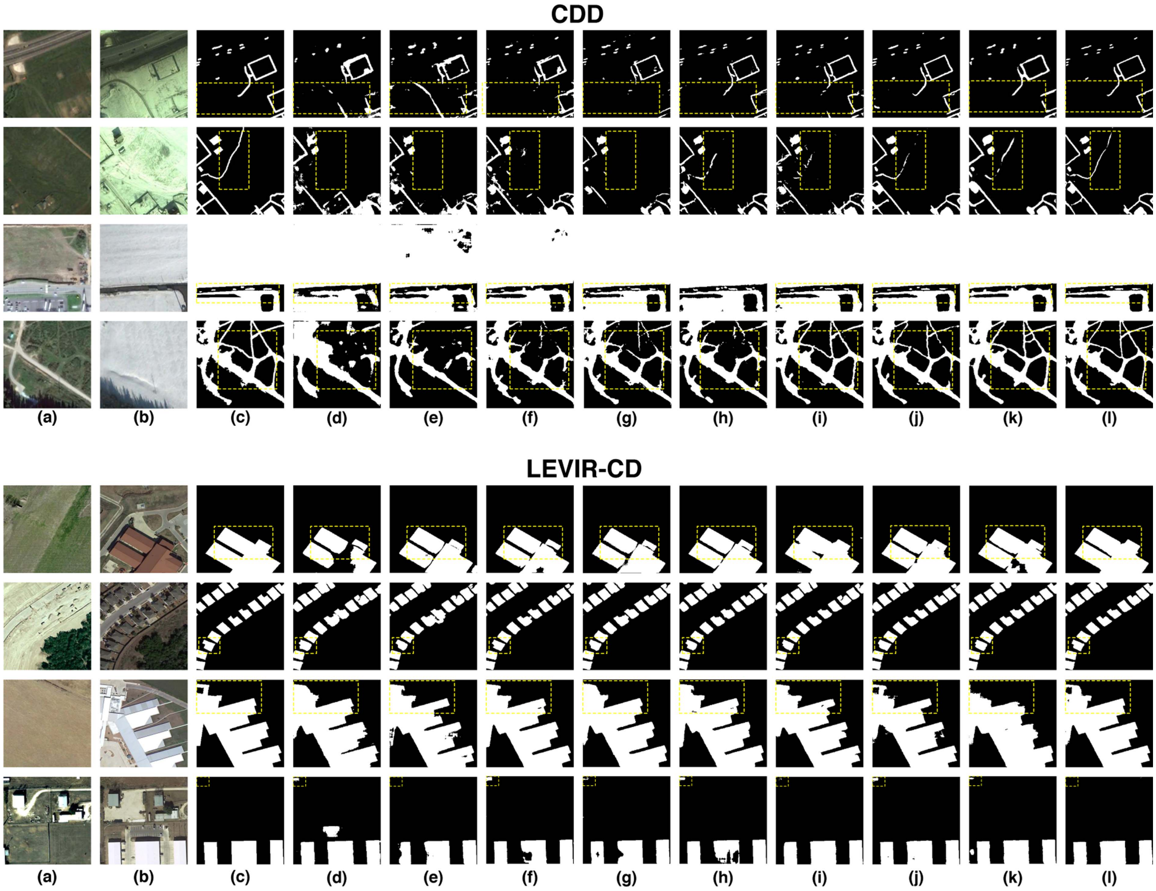

Fig. 4. Visualization comparison of results on the CDD and LEVIR-CD test sets. White is true positive and black is true negative. (a) Pretemporal image. (b) Posttemporal image. (c) Ground truth. (d) FC-Siam-Conc. (e) STANet. (f) FDCNN. (g) DASNet. (h) SNUNet. (i) BIT. (j) HMLNet. (k) USSFC-Net. (l) Our HES-Net.

TABLE III
Quantitative Analysis of Ablation Experiments on the CDD Dataset

Methods	HASM	DFEM	Pre(%)	Rec(%)	F1(%)	Prams(M)
Baseline			95.17	93.82	94.48	12.03
Baseline	√		96.84	96.62	96.73	12.03
Baseline		√	96.64	96.16	96.40	12.07
Baseline	√	√	**97.67**	**96.77**	**97.22**	12.07

The bold values represent the optimal performance for each indicator.

TABLE IV
Quantitative Analysis of Ablation Experiments on the LEVIR-CD Dataset

Methods	HASM	DFEM	Pre(%)	Rec(%)	F1(%)	Prams(M)
Baseline			84.75	89.60	87.11	12.03
Baseline	√		**90.79**	89.19	89.99	12.03
Baseline		√	87.56	89.09	88.04	12.07
Baseline	√	√	90.74	**90.33**	**90.53**	12.07

The bold values represent the optimal performance for each indicator.

a combination of Siamese network and U-Net (Siam-UNet). Introducing HSAM, which embeds hyperbolic information, leads to an improvement of 2.25% and 2.88% in F1 on the CDD and LEVIR-CD datasets, respectively. It is worth noting that the application of HSAM does not introduce any additional parameters to the network. This is because the process of mapping Euclidean space to hyperbolic space and computing HD is solely dependent on the positions and does not require training or optimization of the pseudohyperbolic space and its related parameters. In addition, applying HSAM only results in a 1.477-GB increase in FLOPs. This is primarily because the size of the deep features where HSAM is applied is relatively small. From the tables, it can be observed that combining DFEM with the baseline results in a less significant improvement in F1 compared to HSAM alone but still achieves a remarkable increase of 1.92% and 0.93% on the CDD and LEVIR-CD datasets, respectively. Furthermore, adding DFEM to the HSAM-based model further enhances the exploitation of hyperbolic information, resulting in a greater improvement of 2.74% and 3.42% in F1 compared to the baseline on the CDD and LEVIR-CD datasets, respectively. This clearly demonstrates the effectiveness of both HSAM and DFEM.

V. Discussion

A. Discussion on the Effectiveness of HSAM and DEFM

HSAM can embed hyperbolic information into the network and deeply integrate the features from bitemporal phases, enhancing the network's capability to extract details and resist noise. Specifically, the introduction of HSAM improves the detection ability for details in small regions, blurry areas, and

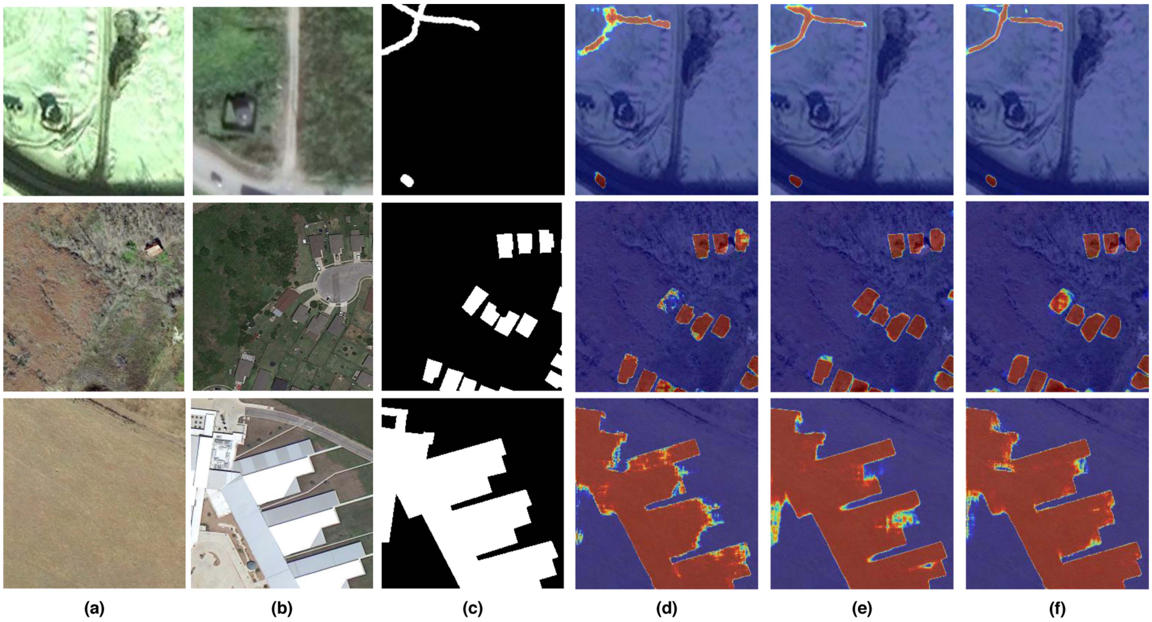

Fig. 5. Feature activation after HSAM or DFEM on the CDD and LEVIR-CD test set. (a) Pretemporal image. (b) Posttemporal image. (c) Ground truth. (d) Feature activation map of baseline. (e) Feature activation map of baseline + HSAM. (f) Feature activation map of baseline + DFEM. Red and yellow in (d)–(f) denote higher attention values.

areas where it is difficult to determine whether there is a change. The feature activation maps of the baseline model and the model with HSAM are shown in Fig. 5(d) and (e), respectively. It can be observed that the model with HSAM effectively complements the missed detections in the original model, such as the buildings in the middle of the second row and the small area on the right side of the middle; as well as the bottom-left corner in the third row. Moreover, there is no apparent noise in the activated maps after introducing HSAM. The model's ability to capture change regions is significantly improved with the introduction of HSAM. Furthermore, the introduction of hyperbolic space can reduce errors caused by factors such as height differences in the images. For example, in the bottom-left corner of the second row, both map (d) and map (f) assign equal attention to both the balcony and the roof in image (b), treating them as the same region of interest. Only map (e) distinguishes this difference, which aligns with the ground truth.

In comparison, the feature activation maps after introducing DFEM [see Fig. 5(f)] show improvements in a different direction. As shown in the first row, although map (e) detects most of the changed areas in the top-left intersection, it fails to refine the edges, while map (f) separates the edges of the intersection. Although map (f) has fewer correctly judged pixels overall compared to map (e), its shape is closer to the ground truth. This is due to the effective utilization of shallow fine-grained features by DFEM. Similarly, in the second and third rows, map (f) exhibits a higher focus on the edges of the change regions compared to map (e). In the gap between the buildings in the middle of the third row, the edges in map (f) better match the shape of the ground truth. The analysis of the data in Tables III and IV, combined with the visual analysis in Fig. 5, thoroughly demonstrates the effectiveness of our proposed HSAM and DFEM.

B. Performance Comparison of HSAM at Different Convolutional Layers

To introduce hyperbolic information into the network and fuse the paired features, HSAM projects a pair of features at the same scale into hyperbolic space and calculates hyperbolic similarity. However, features at different scales possess distinct characteristics, and the effectiveness of HSAM varies significantly when applied to different convolutional layers of the network. Our HES-Net consists of four convolutional layers in the encoder, providing us with four scales of features. The application of HSAM to each layer's features is shown in Fig. 6. For clarity, we refer to the method of applying HSAM after the xth layer convolution as "x H-Layer." Since the receptive field of the first convolutional layer is small and lacks tight contextual information, it is not suitable for applying the hyperbolic attention module. Therefore, our experiments start from the 2 H-Layer, gradually exploring the impact of the HSAM position on model accuracy. To maintain consistency, all experiments default to using DFEM.

From Fig. 6, it is evident that the changes in the position of HSAM have a similar trend on the CDD and LEVIR-CD datasets. The introduction of HSAM at the 2 H-Layer significantly decreases the F1 score compared to the baseline. Interestingly, even when HSAM is applied to multiple convolutional layers, introducing it at the second layer reduces the network's accuracy compared to only applying HSAM at the (3,4) H-Layer. This may be because the features extracted by the second convolutional layer are relatively shallow and cannot fully exhibit the hyperbolic properties we need. Additionally, the introduction of HSAM integrates the noise from the shallow features, amplifying the impact of noise and interfering with network learning.

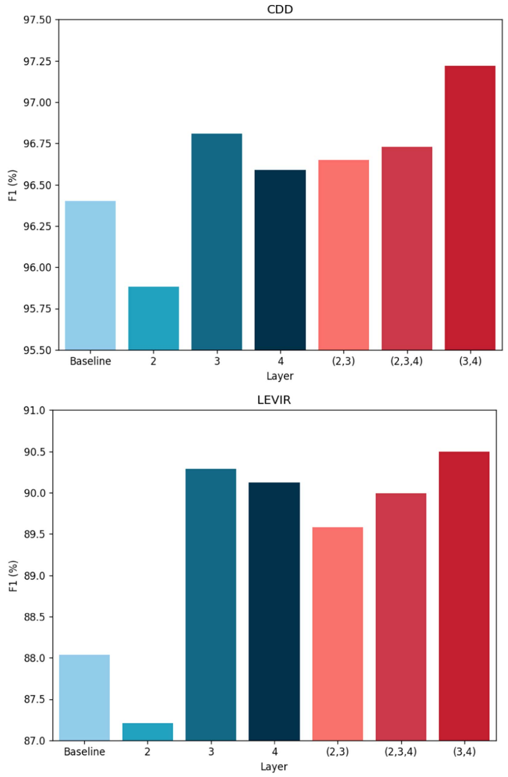

Fig. 6. Effect of HSAM applied at different layers on network accuracy. The horizontal axis represents the applied convolutional layers (*x*) for HSAM, and the vertical axis represents the F1 score of the corresponding method.

The 3 H-Layer and 4 H-Layer significantly improve the network's performance, but the 4 H-Layer falls slightly behind. The deep-level features extracted by the fourth layer possess stronger semantic information but have a poor perception of details, making it difficult to fully utilize HSAMs ability to recognize small objects and detailed information. The features from the third layer combine some advantages of shallow and deep-level features. Data analysis shows that the third layer's features are most suitable for HSAM while the fourth layer's features also reflect the impact of HSAM.

To leverage the distinctive advantages of multiscale features across different layers, we apply HSAM at multiple levels in the encoder. We ultimately select the (3,4) H-Layer version, which yields the best performance. In this version, HSAM effectively extracts hyperbolic information from deep-level features, avoids the influence of noise in shallow features, and obtains more fine-grained features. This version outperforms the suboptimal 3 H-Layer version by 0.41$\%$ on CDD and 0.24$\%$ on LEVIR-CD.

In conclusion, our HSAM demonstrates advantages in the CD task but needs to be applied to deeper layers with multiscale features. Features that are too shallow severely impact the performance of HSAM while features that are too deep fail to fully exploit the benefits of HSAM.

C. Limitations and Future Work

The proposed HES-Net in this article has achieved more accurate results in CD of RS images by introducing hyperbolic information. This demonstrates the effectiveness of hyperbolic space in feature extraction and deep fusion of RS images. However, during the experiments, we have identified certain limitations of the proposed method.

First, in order to demonstrate the excellent performance of HSAM on the basic backbone, we selected the Siamese and U-Net architecture, which is the most fundamental architecture for CD tasks in recent years, as the backbone network. However, there have been many stronger backbone networks emerging now, such as transformers and their various variants [20], [21], as well as improved Siamese networks [27], [41], [42]. These stronger backbone networks can provide more accurate discriminative features, which would be beneficial for extracting hyperbolic information.

Second, since our method only extracts hyperbolic information in deeper convolutional layers, the newly added hyperbolic components in the features lack fine-grained information. This may slightly reduce the precision of the detected boundaries in the new regions.

On the other hand, to further explore the application of hyperbolic information in CD tasks, we plan to extract and fuse hyperbolic information on more powerful backbone networks in the future. Thanks to the plug-and-play nature of HSAM, it is feasible to transfer the module to other networks. Additionally, hyperspectral imagery collected from airborne or satellite sources inevitably suffers from spectral variability [55]. However, the noise resistance brought about by the negative curvature and nonlinear properties of hyperbolic space may help alleviate the impact of spectral variations. Furthermore, HSAM operates on feature maps rather than the images themselves. These advantages provide possibilities for handling multiple input types, such as hyperspectral images and synthetic aperture radar images. In the future, we intend to investigate the performance of hyperbolic information on multimodal data by improving the backbone network.

VI. Conclusion

In this article, we propose a network called HES-Net for RS image CD. The network backbone adopts the popular Siamese-UNet structure. By introducing our designed HSAM, we embed hyperbolic information, which is difficult to extract from RS images, into feature learning. HSAM constructs bitemporal attention based on hyperbolic similarity to fully integrate the features from the two temporal phases and enhance the compactness of change information. Through extensive experimental analysis, we demonstrate the importance of hyperbolic information in the CD.

Furthermore, to address the issue of neglecting shallow fine-grained features and the underutilization of deep-level hyperbolic information, we propose the DFEM. DFEM highlights the advantages of multiscale features by deep-level feature subtraction, simultaneously enhancing HSAM and strengthening fine-grained features to refine the boundaries of change regions.

The effectiveness of these methods is validated through experiments on two popular datasets: 1) CDD and 2) LEVIR-CD. The experimental results demonstrate that our HES-Net outperforms existing popular networks in terms of detection accuracy.

References

[1] P. R. Shekar and A. Mathew, "Detection of land use/land cover changes in a watershed: A case study of the Murredu watershed in Telangana State, India," *Watershed Ecol. Environ.*, vol. 5, pp. 46–55, 2023.

[2] C. Mucher, K. Steinnocher, F. Kressler, and C. Heunks, "Land cover characterization and change detection for environmental monitoring of pan-Europe," *Int. J. Remote Sens.*, vol. 21, no. 6/7, pp. 1159–1181, 2000.

[3] A. Novo-Fernández, S. Franks, C. Wehenkel, P. M. López-Serrano, M. Molinier, and C. A. López-Sánchez, "Landsat time series analysis for temperate forest cover change detection in the Sierra Madre Occidental, Durango, Mexico," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 73, pp. 230–244, 2018.

[4] 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.

[5] Z. Zheng, Y. Zhong, J. Wang, A. Ma, and L. Zhang, "Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters," *Remote Sens. Environ.*, vol. 265, 2021, Art. no. 112636.

[6] L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 3, pp. 1171–1182, May 2000.

[7] D. Hong et al., "SpectralGPT: Spectral foundation model," 2023, *arXiv:2311.07113*.

[8] D. Hong, J. Yao, C. Li, D. Meng, N. Yokoya, and J. Chanussot, "Decoupled-and-coupled networks: Self-supervised hyperspectral image super-resolution with subpixel fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Oct. 2023, Art. no. 5527812.

[9] J. Lin, F. Gao, X. Shi, J. Dong, and Q. Du, "SS-MAE: Spatial–spectral masked autoencoder for multisource remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Nov. 2023, Art. no. 5531614.

[10] M. Wang, F. Gao, J. Dong, H.-C. Li, and Q. Du, "Nearest neighbor-based contrastive learning for hyperspectral and LiDAR data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Mar. 2023, Art. no. 5501816.

[11] A.-R. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 14–22, Jan. 2012.

[12] H. Chen, C. Wu, B. Du, L. Zhang, and L. Wang, "Change detection in multisource VHR images via deep Siamese convolutional multiple-layers recurrent neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2848–2864, Apr. 2020.

[13] X. Li et al., "MAM-RNN: Multi-level attention model based RNN for video captioning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 2208–2214.

[14] B. Zhao, H. Li, X. Lu, and X. Li, "Reconstructive sequence-graph network for video summarization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2793–2801, May 2022.

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2015, pp. 234–241.

[17] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 539–546.

[18] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected Siamese network for change detection of VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Feb. 2022, Art. no. 8007805.

[19] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional Siamese networks for change detection," in *Proc. IEEE 25th Int. Conf. Image Process.*, 2018, pp. 4063–4067.

[20] W. G. C. Bandara and V. M. Patel, "A transformer-based Siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 207–210.

[21] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5920416.

[22] Y. Feng, H. Xu, J. Jiang, H. Liu, and J. Zheng, "ICIF-Net: Intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 4410213.

[23] C. Zhang, L. Wang, S. Cheng, and Y. Li, "SwinSUNet: Pure transformer network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art. no. 5224713.

[24] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[25] X. Li et al., "Hybridizing Euclidean and hyperbolic similarities for attentively refining representations in semantic segmentation of remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Nov. 2022, Art. no. 5003605.

[26] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: Going beyond Euclidean data," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 18–42, Jul. 2017.

[27] T. Lei et al., "Difference enhancement and spatial–spectral nonlocal network for change detection in VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Dec. 2022, Art. no. 4507013.

[28] H. Chen, F. Pu, R. Yang, R. Tang, and X. Xu, "RDP-Net: Region detail preserving network for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Dec. 2022, Art. no. 5635010.

[29] C.-P. Chen, J.-W. Hsieh, P.-Y. Chen, Y.-K. Hsieh, and B.-S. Wang, "SARAS-Net: Scale and relation aware Siamese network for change detection," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 14187–14195.

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

[31] Y. Shu, W. Li, M. Yang, P. Cheng, and S. Han, "Patch-based change detection method for SAR images with label updating strategy," *Remote Sens.*, vol. 13, no. 7, 2021, Art. no. 1236.

[32] Y. Han, A. Javed, S. Jung, and S. Liu, "Object-based change detection of very high resolution images by fusing pixel-based change detection results using weighted Dempster–Shafer theory," *Remote Sens.*, vol. 12, no. 6, 2020, Art. no. 983.

[33] 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 Sens.*, vol. 12, no. 2, 2020, Art. no. 205.

[34] T. Liu, L. Yang, and D. Lunga, "Change detection using deep learning approach with object-based image analysis," *Remote Sens. Environ.*, vol. 256, 2021, Art. no. 112308.

[35] Y. Bin, M. Yin, C. Jin, L. Jianqiang, C. Jie, and Y. Kai, "Review of remote sensing change detection in deep learning: Bibliometric and analysis," *Nat. Remote Sens. Bull.*, vol. 27, no. 9, pp. 1988–2005, 2023.

[36] R. Liu, M. Kuffer, and C. Persello, "The temporal dynamics of slums employing a CNN-based change detection approach," *Remote Sens.*, vol. 11, no. 23, 2019, Art. no. 2844.

[37] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Multitask learning for large-scale semantic change detection," *Comput. Vis. Image Understanding*, vol. 187, 2019, Art. no. 102783.

[38] H. Lyu and H. Lu, "Learning a transferable change detection method by recurrent neural network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 5157–5160.

[39] M. Rußwurm and M. Korner, "Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 1496–1504.

[40] D. Peng, Y. Zhang, and H. Guan, "End-to-end change detection for high resolution satellite images using improved UNet," *Remote Sens.*, vol. 11, no. 11, 2019, Art. no. 1382.

[41] T. Lei et al., "Ultralightweight spatial–spectral feature cooperation network for change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Mar. 2023, Art. no. 4402114.

[42] J. Chen et al., "DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1194–1206, Nov. 2021.

[43] 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.

[44] W. Peng, T. Varanka, A. Mostafa, H. Shi, and G. Zhao, "Hyperbolic deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 10023–10044, Dec. 2022.

[45] O. Ganea, G. Bécigneul, and T. Hofmann, "Hyperbolic neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 5350–5360.

[46] V. Khrulkov, L. Mirvakhabova, E. Ustinova, I. Oseledets, and V. Lempitsky, "Hyperbolic image embeddings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6417–6427.

[47] R. P. Duin and E. Pkalska, "Non-Euclidean dissimilarities: Causes and informativeness," in *Proc. Structural, Syntactic, Stat. Pattern Recognit.: Joint IAPR Int. Workshop*, 2010, pp. 324–333.

[48] M. Yang et al., "Hyperbolic graph neural networks: A review of methods and applications," 2022, *arXiv:2202.13852*.

[49] Y. Zhang, H. Zhu, J. Liu, P. Koniusz, and I. King, "Alignment and outer shell isotropy for hyperbolic graph contrastive learning," 2023, *arXiv:2310.18209*.

[50] J. W. Cannon et al., "Hyperbolic geometry," *Flavors Geometry*, vol. 31, no. 59–115, 1997, Art. no. 2.

[51] M. Lebedev, Y. V. Vizilter, O. Vygolov, V. A. Knyaz, and A. Y. Rubis, "Change detection in remote sensing images using conditional adversarial networks," *Int. Arch. Photogrammetry, Remote Sens., Spatial Inf. Sci.*, vol. 42, pp. 565–571, 2018.

[52] 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.

[53] M. Zhang and W. Shi, "A feature difference convolutional neural network-based change detection method," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7232–7246, Oct. 2020.

[54] Y. Liang, C. Zhang, J. Liu, and M. Han, "HMLNet: A hierarchical metric learning network with dual attention for change detection in high-resolution remote sensing images," *Int. J. Remote Sens.*, vol. 44, no. 3, pp. 1001–1021, 2023.

[55] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

Qian Yang received the B.E. degree in information engineering in 2022 from the Qingdao University of Science and Technology, Qingdao, China, where he is currently working toward the master's degree in electronic information technology from Qingdao University of Science and Technology.

His research interests include computer vision, change detection, and remote sensing.

Shujun Zhang received the Ph.D. degree in artificial intelligence in virtual marine environment from the Ocean University of China, Qingdao, China, in 2007.

She was a Postdoctoral Researcher with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing. She is currently an Associate Professor with the Qingdao University of Science and Technology, Qingdao. She is also a Communication Evaluation Expert with China Academic Degrees and Graduate Education Development Center, Beijing. Her research interests include computer vision, image processing, machine learning, and virtual reality.

Dr. Zhang is a member of the Chinese Computer Federation.

Jinsong Li received the B.E. degree in computer science and technology from Shandong University, Jinan, China, in 2021. He is currently working toward the master's degree in computer technology with the School of Data Science, Qingdao University of Science and Technology, Qingdao, China.

His research interests include computer vision and remote sensing.

Yukang Sun is currently working toward the master's degree in software engineering with the School of Data Science, Qingdao University of Science and Technology, Qingdao, China.

His research interests include computer vision, image processing, and deep learning.

Qi Han is currently working toward the master's degree in computer technology with the School of Data Science, Qingdao University of Science and Technology, Qingdao, China.

His research interests include computer vision, image processing, and deep learning.

Yuanyuan Sun received the Ph.D. degree in remote sensing information technology from Zhejiang University, Hangzhou, China, in 2018.

She is currently a Lecturer with the Qingdao University of Science and Technology, Qingdao, China. Her research interests include applied remote sensing, computer vision, pattern recognition, and data mining.