Bitemporal Attention Sharing Network for Remote Sensing Image Change Detection

Zhongchen Wang, Guowei Gu, Min Xia, *Member, IEEE*, Liguo Weng, and Kai Hu

***Abstract*—With the advancement of remote sensing image technology, the availability of very high-resolution image data has brought new challenges to change detection (CD). Currently, deep learning-based CD methods commonly employ bitemporal interaction networks using convolutional neural networks or transformers. Yet, these models overly emphasize object accuracy, leading to a significant increase in computational costs with limited performance gains. In addition, the current bitemporal interaction mechanisms are simplistic, failing to adequately account for spatial positions and scale variations of different objects, resulting in an inaccurate modeling of dynamic feature changes between images. To address these issues, a bitemporal attention sharing network is proposed, which tackles the problems effectively by making bitemporal and multiscale attention sharing the primary mode of feature interaction. Specifically, the proposed bitemporal attention sharing module leverages pairs of features preliminarily encoded by a backbone to construct shared global features, directing attention to target changes. Then, through cross-scale attention guidance and weighted fusion, it achieves attention sharing of multiscale features, eliminating the need for overrelying on deep convolutional layers for feature extraction. Experiments on three public datasets demonstrate that, in comparison to several state-of-the-art methods, our model achieves superior performance with low computational cost.**

***Index Terms*—Attention mechanism, convolutional neural network (CNN), change detection (CD), remote sensing (RS) images, transformer.**

I. Introduction

CHANGE detection (CD) is a crucial task in the field of remote sensing, aimed at identifying changes between images acquired at different times within the same geographical area. The objective of CD is to detect target changes, particularly those related to human activities, such as environmental or land use changes, while avoiding background changes caused by seasonal variations, shadows, atmospheric conditions, and lighting changes.

Recently, significant advancements have been made in remote sensing technology, driven by the development of various sensors including airborne sensors and satellites, such as the Landsat series and Gaofen series [1], [2], [3], [4]. The advancements in data acquisition techniques have made it possible to accurately obtain multiple images of the same geographical area at different times. This has led to a revolutionary shift in CD, enabling the identification and localization of Earth surface changes with the aid of VHR imagery [5]. Supported by remote sensing technology, CD plays a vital role in various applications, such as damage assessment [6], urban planning [7], ecosystem monitoring [8], and resource management [9]. It has become a research hotspot, driving innovation of Earth observation.

In the early stages of CD development, due to limitations in the hardware development of satellites, optical instruments, and other factors, the quality of remote sensing data was poor, characterized by low resolution, making it challenging to support VHR detection. Simple algebraic calculations or direct pixel comparisons were commonly employed as cost-effective detection methods. Both principal component analysis [10] and change vector analysis (CVA) [11] are pixel-based CD methods. PCV performs direct pixel value comparison, whereas CVA represents pixels as vectors and calculates vector differences. This enables CVA to handle more complex images and scenes effectively. Both methods require image preprocessing to correct distortions caused by background changes, and the results are typically binarized using thresholding or clustering techniques [12].

With the rise of machine learning algorithms, such as support vector machines [13] and decision trees [14], alongside advancements in remote sensing image technology, object-level CD methods emerged. By considering global information, they reduce the salt and pepper effect and achieve better modeling of contextual information within images while also accounting for relationships between adjacent pixels [15]. These techniques utilize approaches such as object segmentation, contour analysis, fusion of photometric and textural differences, and statistical testing. They extract object features from multitemporal image pairs to identify changes in object status [16]. These methods overcome the reliance on manual intervention and the poorer robustness exhibited by pixel-level methods. However, despite these advancements, machine learning-based CD methods still face two challenging issues. First, they are susceptible to limitations posed by data volume and sample distribution [17]. CD tasks commonly encounter class imbalances and uneven sample distributions, potentially leading to insufficient data or sample bias issues in machine learning methods. Second, there is the challenge of modeling complex relationships. Some intricate change patterns and relationships may require more powerful modeling capabilities to capture and represent. Machine learning

Manuscript received 17 January 2024; revised 5 April 2024; accepted 11 May 2024. Date of publication 14 May 2024; date of current version 30 May 2024. This work was supported by the National Natural Science Foundation of China under Grant 42075130. *(Correspondence author: Min Xia.)*

The authors are with the Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: xiamin@nuist.edu.cn).

Our code is available at https://github.com/JasonWZC/BASNet

Digital Object Identifier 10.1109/JSTARS.2024.3400925

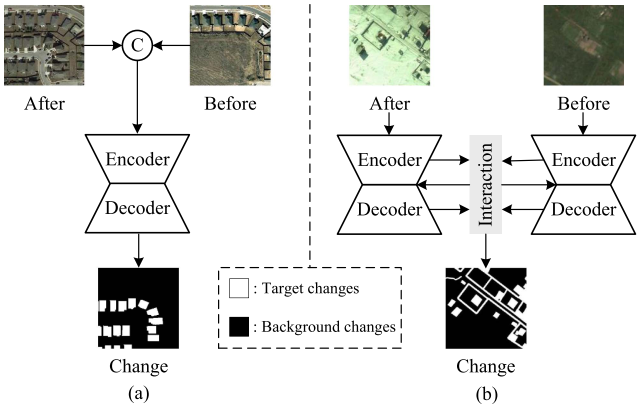

Fig. 1. Two common CD frameworks. (a) Single-stream. (b) Double-stream.

methods might struggle to fully utilize hidden patterns and complex relationships in large-scale data, thus limiting performance.

Fortunately, in recent years, the latest advancements in deep learning within computer vision have provided many promising solutions for CD problems [18]. Two mainstream CD design frameworks have emerged, known as the single-stream framework and the double-stream framework [19], as shown in Fig. 1. Considering that CD is a downstream task of semantic segmentation, the single-stream framework leverages mature semantic segmentation algorithms, such as [20] and [21]. They involve inputting bitemporal images into the same single-temporal semantic segmentation model and obtaining global feature maps either through simple fusion at the initial input or final output stage, or by subtracting to obtain difference feature maps. While this framework is plug-and-play, it has two main drawbacks. First, the multiscale features generated by the encoder–decoder are challenging to correlate, leading to limited interaction between the bitemporal features and making the model susceptible to background changes interference [22], [23]. Second, since semantic segmentation is designed for single-temporal tasks, it often uses high depth to improve segmentation accuracy for small objects and edges [24], [25]. When directly applied to CD, this approach would double the computational cost unnecessarily. Thus, the double-stream framework, which focuses on bitemporal interaction, is a more suitable choice. Through thoughtful design, this framework can eliminate background changes with the same semantic concept based on different spectral features of bitemporal images, while also exploring target changes related to different semantic concepts [26].

Despite the current development of deep learning-based CD methods toward the double-stream framework, two significant challenges still persist. First, intentional bitemporal feature fusion or exchange can lead to confusion in semantic information between the bitemporal images, as seen in [27] and [28]. Networks that frequently adopt pixel-level differencing, addition, channel concatenation, or direct feature swapping may reduce the ability to perceive background changes effectively. Second, as seen in [29] and [30], an excessive reliance on deeper backbone networks and multiple convolutional layer modules has emerged, resulting in significant computational costs for marginal performance improvements. This scenario has raised the threshold for the practical application of CD.

Some recent works attempt to address these two types of challenges. For instance, SARASNet [31] employs a relation-aware module to analyze deep-learned multichannel features before performing subtraction on bitemporal features. This enhances the interaction between features extracted from two input images, helping the model efficiently handle data with highly imbalanced samples. However, SARASNet suffers from a large number of parameters. Another example is DPCCNet [32], which utilizes an improved lightweight ResNet50 as the backbone network and employs dual-perspective fusion (DPF) to explore temporal information between paired images. DPF uses one of the bitemporal images as a reference and queries the reference image with the other image to identify differences. Since the selection of the reference image does not interfere with the recognition of changing regions, the model can consider change information from two perspectives [33]. This is advantageous for the model to cope with complex change scenarios. However, the overall accuracy is constrained by the lightweight backbone. Therefore, achieving efficient interaction while balancing computational cost and accuracy is a key focus of research. Hence, we are dedicated to developing a double-stream structure for a bitemporal interaction network, emphasizing the construction of same-scale and cross-scale attention-sharing mechanisms to replace the usage of certain convolutional layers, effectively improving computational efficiency. Specifically, the network first utilizes a pruned ResNet18 as the backbone network for feature encoding. Then, multiscale features from two temporal states construct a shared attention through the bitemporal attention sharing (BAS) module. The attention is guided to target changing regions, addressing the issue of semantic information confusion in bitemporal features. The cross-scale attention guidance (CAG) module, connecting high-level and low-level features, enhancing the efficiency of utilizing stage features to reduce reliance on the backbone network. Next, pixel-level addition and subtraction are employed to explore homogeneous and differential information in bitemporal features, respectively. Finally, the weighted fusion (WF) module is used for the decoding of the dual features.

Our main contributions in this work are as follows.

1) A BAS module is proposed, which can capture the global attention of bitemporal images to build a bitemporal shared attention sequence. This module guides attention toward target changes and suppresses attention to background changes, achieving efficient bitemporal information interaction.
2) Two cross-scale attention sharing modules, namely CAG and WF, are proposed. CAG leverages attention parameters from high-level features to weight low-level features, optimizing the attention density of output features at different stages and enhancing the utilization efficiency of features. WF can gather global attention from single temporal multiscale images to construct multiscale shared attention. It then applies this shared attention to weight multiscale features, suppressing semantic information loss in the decoding stage.

3) Quantitative and qualitative experiments on three public datasets demonstrate that our proposed BASNet outperforms the state-of-the-art methods in the field of CD.

II. Related Work

A. CD Methods Based on CNN

With the remarkable advancements of CNNs in computer vision, a plethora of CNN-based CD methods have emerged. Daudt et al. [27] introduced the FC-CD series based on fully convolutional neural networks, enabling end-to-end pixel-level CD. This approach utilizes skip connections to learn contextual information from images and exhibits exceptionally fast detection speed. However, it struggles to handle highly complex scenarios. ChangeNet [34] employs multiscale features, excels in object tracking, and finds utility in domains such as video surveillance and action recognition. IFNet [35] and DTCDSCN [36] employ channel or spatial attention mechanisms, to focus on significant change areas, yet they face challenges distinguishing small objects from noise. Changer [37] and SAGNet [38] intertwine bitemporal interaction schemes between encoding levels, combining hybrid layers and backbone structures to enhance the similarity in bitemporal feature distribution. This integration achieves automatic domain adaptation between bitemporal domains.

B. CD Methods Based on Transformer and Self Attention (SA)

The widespread adoption of transformer-based methods in natural language processing [39] has provided new insights for computer vision [40]. Many researchers have endeavored to design networks based purely on the transformer backbone [41], [42], [43]. ChangeFormer [44] unifies layered transformer encoders with the multilayer perceptron decoders of siamese network architecture, effectively rendering multiscale distant details. SwinSUNet [45] solely employs a hierarchical swin transformer as the encoder–decoder, allowing SA computations within segmented windows and utilizing window shifts to associate global information. However, the pure transformer incurs substantial computational costs and often requires larger datasets compared to CNNs to realize its potential due to the lack of inductive biases.

The fusion of CNN with transformer has emerged as a more economical option [46]. STANet [47] combines multiscale pooling and SA mechanisms to model spatial–temporal relationships, extracting more discriminative features. BIT [48] initially employs a CNN backbone to extract features from bitemporal images and transforms them into token sequences. It then uses a transformer encoder–decoder to connect token-based contextual information and feedbacks enriched tokens to the pixel space, refining original features. ACABFNet [49] connects CNN and transformer backbones in parallel to bridge their hierarchical features and employs cross-attention for bitemporal interaction. DPCCNet [32] enhances temporal information extraction through the fusion of bitemporal data and context modeling.

III. Methodology

As shown in Fig. 2, the architecture of BASNet comprises two main components. The first part involves feature encoding, executed by the backbone network. To minimize computation costs, we select the lightweight ResNet18 as the backbone and remove its last layer. The bitemporal images undergo hierarchical encoding to generate three distinct hierarchical features of different sizes. The second part involves attention sharing, categorized into same-scale and cross-scale attention sharing. Same-scale attention sharing includes BAS modules, elementwise addition and subtraction. BAS utilizes features from both temporal images to construct shared global attention tokens, enhancing their attention distributions. Elementwise addition captures common information and shared features between the two moments. Conversely, elementwise subtraction extracts differences between the two temporal instances, allowing the model to focus more on areas with actual changes. Cross-scale attention sharing encompasses CAG module and WF module. CAG employs attention from high-level features to guide low-level features, enhancing their attention intensity. WF harmoniously fuses interaction features, laying the foundation for generating prediction maps.

A. BAS Module

Differences naturally exist in remote sensing images captured at different time points, yet only the changes in the objects of interest demand attention. Effectively interacting bitemporal features to suppress interference from background changes and uncover regions of shared attention represents a critical challenge [31], [50]. To address this, we have devised BAS module, as shown in Fig. 3 which consists of two parts. First, a pair of transformer blocks based on SA mechanism and feed-forward network are employed. Second, a global attention sharing layer is sandwiched between these two blocks. Initially, after passing through three convolutions, the input features are subjected to channel transformations to map them into the feature space of query (Q), key (K), and value (V). This facilitates subsequent SA computations. To be more precise, the triple features resulting from channel mappings are denoted as $F_q \in \mathbb{R}^{\frac{C}{r}\times H\times W}$, $F_k \in \mathbb{R}^{\frac{C}{r}\times H\times W}$, and $F_v \in \mathbb{R}^{C\times H\times W}$. In order to align with transformer calculations, the reduction factor r for both F_q and F_k should remain consistent. This operation not only aids in amalgamating channel information but also reduces computational complexity. Each pixel of the triple features serves as a token, and subsequently, these tokens are flattened into a sequence $Q, K, V \in \mathbb{R}^{C\times N}$ that can be understood by the transformer block. Here H and W, respectively, signify the vertical and horizontal pixel count of the feature map, C denotes the channel count of the feature map and tokens, and $N = HW$ reflects the total number of pixels in the transformed sequence of tokens. The triple sequence can be represented as follows:

$$\begin{aligned}
Q &= \text{Reshape}\left(\text{Conv}_q(F_q)\right)\\
K &= \text{Reshape}\left(\text{Conv}_k(F_k)\right)\\
V &= \text{Reshape}\left(\text{Conv}_v(F_v)\right)
\end{aligned} \tag{1}$$

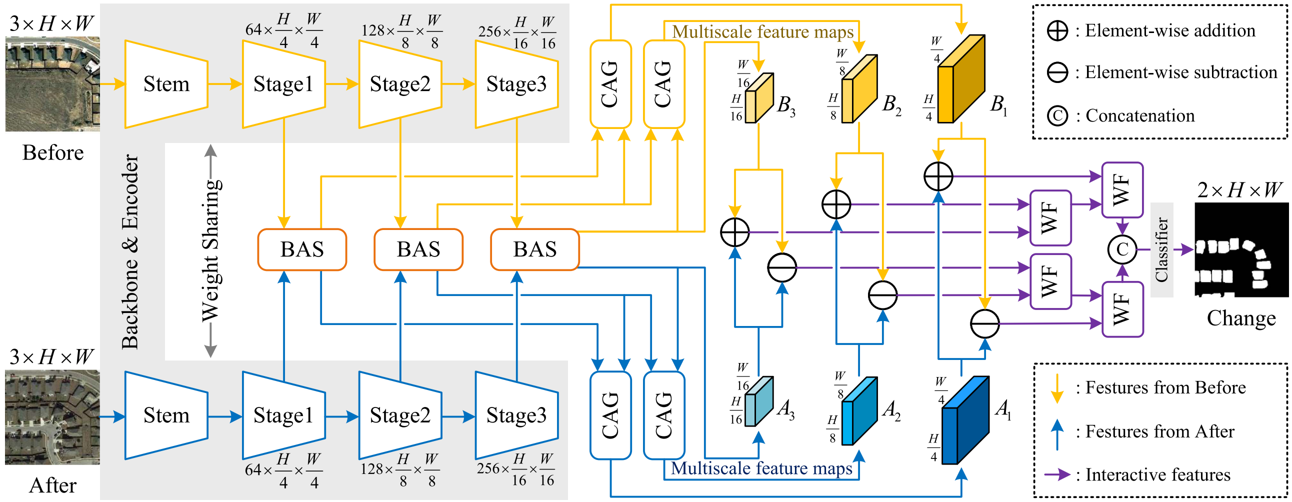

Fig. 2. Overall structure of BASNet.

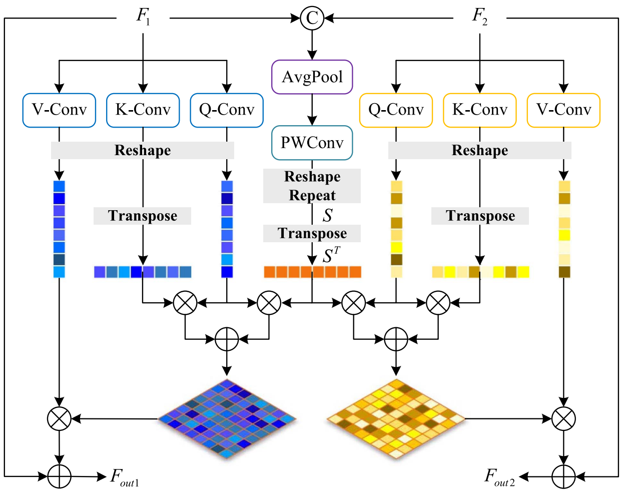

Fig. 3. Structure of BAS module.

where Reshape$(\cdot)$ denotes the transformation of a matrix into sequence tokens. $\mathrm{Conv}_{\mathrm{q,k,v}}(\cdot)$ denote three types of 1×1 convolution operations.

Then, Q and the transposed K calculate their similarity through matrix multiplication and provide feedback to V. The traditional transformer SA is computed using the following dot product formula:

$$\mathrm{Att}(Q, K, V) = \delta\left(Q \otimes K^T\right) \otimes V \tag{2}$$

where $\delta(\cdot)$ denotes softmax activation function, T indicates a transpose operation, and $\otimes$ represents matrix multiplication.

To enhance the interaction between a pair of transformer blocks, we concatenate the input bitemporal features along the channel dimension. We then employ adaptive average pooling to compress the feature size to 1×1 and reduce the number of channels to achieve global channel attention. In order to obtain similar information mapping capability as K, single-pixel feature matrix is treated as an individual token and cloned to itself, resulting in a total of N pixels for the sequence tokens. This sequence is denoted as the attention sharing sequence $S \in \mathbb{R}^{C\times N}$. The calculation formula for the above-mentioned process can be expressed as

$$S = \mathrm{Repeat}(\mathrm{Reshape}(\mathrm{PWConv}(\mathrm{Avgpool}(\mathrm{Cat}[F_1, F_2])))) \tag{3}$$

where Repeat$(\cdot)$ denotes unit replication. PWConv$(\cdot)$ refers to pointwise convolution operation. Avgpool$(\cdot)$ indicates adaptive average pooling. Cat$[\cdot]$ represents channel concatenation.

Similarly to the K, S is also transposed and multiplied with the Q to obtain their similarity through matrix multiplication. The two types of similarities are fused by elementwise addition, resulting in the attention sharing matrix. This process allows the global attention to influence the semantic understanding of the single temporal state. After normalization using the softmax function, attention weights are obtained and multiplied with the V through matrix multiplication. Finally, the original input feature size is restored through transformation. The attention sharing can be computed using the following dot product formula:

$$\mathrm{AttShare}(Q, K, V, S) = \delta\left(Q \otimes K^T + Q \otimes S^T\right) \otimes V. \tag{4}$$

Finally, a residual connection is applied before the output to prevent potential feature divergence. The bitemporal image will experience the above-mentioned operations synchronously. In terms of functionality, BAS comprehensively learns global information from the single temporal images and effectively guides the attention distribution of the bitemporal images through shared attention. This process suppresses irrelevant interferences and enables more precise identification of regions with changes. This attention-guided interaction modeling, where attention weights guide the interaction, avoids direct contact between the bitemporal features. It reduces semantic confusion caused by traditional interaction modules, such as pixelwise subtraction or cross-attention.

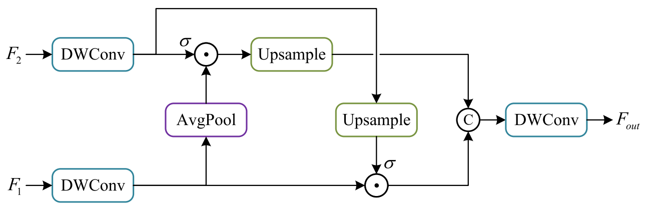

Fig. 4. Structure of CAG module.

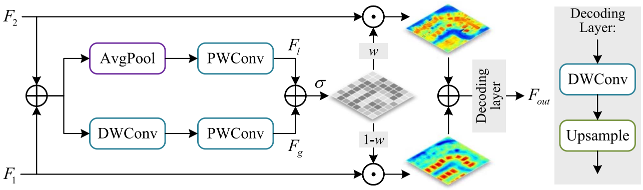

Fig. 5. Structure of WF module.

B. CAG Module

Because our model pursues low computational complexity, the encoder has few convolutional layers, which limits the semantic perception ability of hierarchical features. To make full use of the existing multiscale features, as shown in Fig. 4, we introduce the CAG module. Low-level features $F_1 \in \mathbb{R}^{C\times H\times W}$ guided by high-level features $F_2 \in \mathbb{R}^{2C\times \frac{H}{2}\times \frac{W}{2}}$ in two ways. The first approach is to keep the high-level features unchanged, reduce the channel and pool the low-level features to the same size as the high-order features, and then use the sigmoid activation function to obtain the weight matrix. The weighted low-level features are finally upsampled to restore the size. The first approach can be expressed as

$$F' = \mathrm{Up}(\mathrm{Avgpool}(\mathrm{DWConv}(F_1) \odot \sigma(\mathrm{DWConv}(F_2)))). \quad (5)$$

The second approach involves maintaining the integrity of low-level features while restoring high-level features to match the dimensions of the low-level features through upsampling. A weight matrix is then obtained using the sigmoid activation function, which is used to weight the low-level features. The second approach can be expressed as

$$F'' = \mathrm{DWConv}(F_1) \odot \sigma(\mathrm{Up}(\mathrm{DWConv}(F_2))) \quad (6)$$

where F_1 and F_2 denote low-level and high-level features, respectively. $\mathrm{DWConv}(\cdot)$ denotes depthwise separable convolution operation. $\mathrm{Avgpool}(\cdot)$ denotes adaptive average pooling. $\mathrm{Up}(\cdot)$ denotes upsampling. $\sigma(\cdot)$ denotes sigmoid activation function. $\odot$ represents elementwise multiplication.

The outputs of the two approaches are merged by channelwise fusion. Through this process, CAG guides the information from high-level features to low-level features, enhancing the semantic expression capability of the low-level features. As a result, the low-level features retain multichannel information while gaining richer semantic information. This effectively enhances the overall feature representation.

C. WF Module

Although there may exist subtle differences in the fusion process of features from different stages, they still contain valuable attention information that is worth sharing [51]. Therefore, as shown in Fig. 5, we introduce the WF module to harmonize the differences between multiscale features and effectively fuse the semantic information at various scales. First, it merges features from adjacent scales and performs feature processing both globally and locally, resulting in features denoted as $F_g \in \mathbb{R}^{C\times 1\times 1}$ and $F_l \in \mathbb{R}^{C\times H\times W}$, respectively. This process aims to enhance the contextual awareness of the fused features. Then, to simplify the representation of complex semantic information, we use the sigmoid activation function to convert abstract feature information into weight matrices w and $1-w$, ranging from 0 to 1. This allows the module to perform soft selection or weighted averaging of input features. The calculation formula for the above-mentioned process can be expressed as

$$F_l = \mathrm{PWConv}(\mathrm{Avgpool}(F_1 + F_2)) \quad (7)$$

$$F_g = \mathrm{PWConv}(\mathrm{DWConv}(F_1 + F_2)) \quad (8)$$

$$w = \sigma(F_l + F_g) \quad (9)$$

$$F_{\mathrm{out}} = (1-w) \odot F_1 + w \odot F_2 \quad (10)$$

where F_1 and F_2 denote adjacent-scale features. $\mathrm{PWConv}(\cdot)$ refers to pointwise convolution operation. $\mathrm{DWConv}(\cdot)$ denotes depthwise separable convolution operation. $\mathrm{Avgpool}(\cdot)$ denotes adaptive average pooling. $\sigma(\cdot)$ denotes sigmoid activation function. $\odot$ represents elementwise multiplication.

Within the feature decoding layer, a depthwise convolution followed by upsampling is employed to decrease the feature channel count by half, while simultaneously doubling the spatial dimensions. This progressive feature decoding process enables the gradual restoration and reconstruction of feature representations, thereby furnishing subsequent processing steps with richer and more informative data. WF offers a lightweight approach to handle multiscale feature fusion, effectively preventing feature divergence and enhancing the precision of detecting edges within regions of target change.

IV. Experiments

A. Datasets

The experiments were conducted on three different public datasets, and the input images were data augmented to increase the diversity of the training data, including horizontal flipping, vertical flipping, and random rotation.

LEVIR-CD [52]: This dataset employs VHR remote sensing images obtained from Google Earth. The target changes include variations in different types of buildings found in cities and villages. The images in this dataset are collected over varying time spans, introducing variations caused by seasonal changes and lighting conditions. It effectively validates the ability of network to focus on target changes.

CDD [53]: This dataset employs multiple pairs of remote sensing images captured in different seasons of the same geographical area using Google Earth, encompasses a diverse range of target changes. These changes encompass various types,

TABLE I
EXPERIMENTAL DATASETS PARAMETERS

Dataset	Size (pixel)	Resolution (m/pixel)	Pixel distribution (pcs)			Image distribution (pcs)		
			Target	Background	Ratio	Train	Validation	Test
LEVIR-CD	256 × 256	0.5	30 913 975	637 028 937	1:20.61	7120	1024	2048
CDD	256 × 256	0.03∼2	134 068 750	914 376 178	1:6.83	10000	3000	3000
GZ-CD	256 × 256	0.55	20 045 119	200 155 821	1:10.01	2504	313	313

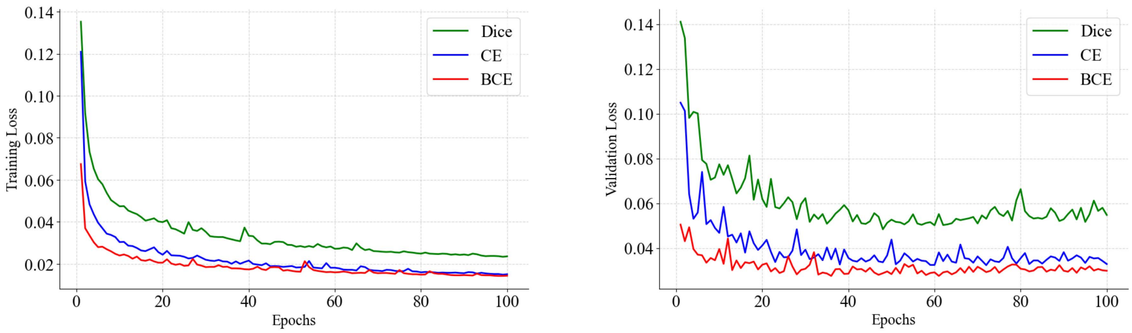

Fig. 6. Training loss curve and validation loss curve generated by BASNet using different loss functions on LEVIR-CD.

including manmade objects of varying sizes, such as roads, cars, buildings, as well as natural objects, such as individual trees and forests. Pronounced seasonal differences result in substantial brightness variations, which make it challenging for the network to distinguish between target changes and background changes.

GZ-CD [54]: This dataset utilizes 19 pairs of remote sensing images acquired from Google Earth, capturing the city of Guangzhou in the years 2006 and 2019. The target changes in this dataset include various types of buildings. Notably, this dataset comprises a smaller number of samples, making it possible to examine the extent to which the network relies on a substantial amount of labeled data through a horizontal comparison with the other two datasets. For further details about the three datasets, refer to Table I.

B. Implementation Details

In terms of hardware, our experiments are conducted using an Intel Core i5-13600KF CPU and an NVIDIA RTX 3090 GPU. On the software side, we employ Python (3.9) and PyTorch (1.13.1). Batch size is set to 32. As depicted in Fig. 6, we evaluate three commonly used loss functions for both semantic segmentation and CD tasks: Dice loss, cross-entropy loss, and BCEWithLogitsLoss. Training is carried out for 100 epochs, and we compare the convergence speed and loss values of these loss functions. Ultimately, we select BCEWithLogitsLoss (BCE) as the optimal choice. It combines the sigmoid activation function and BCE loss to make the calculation more stable and efficient. The mathematical expression is as follows:

$$\mathrm{BCE}(x, y) = -y\log(\sigma(x)) - (1-y)\log(1-\sigma(x)) \tag{11}$$

where x is the model output, y is the real label, $\sigma(\cdot)$ is the sigmoid activation function. We employ the Adam [55] optimizer and utilized the poly learning rate adjustment during network training. The initial learning rate ($\mathrm{Ir}_{\mathrm{base}}$) is set to 0.001, the maximum training iterations (max_epoch) are set to 200, and the learning rate (lr) of each epoch is

$$\mathrm{lr} = \mathrm{lr}_{\mathrm{base}} \times \left(1 - \frac{\mathrm{epoch}}{\mathrm{max\_epoch}}\right). \tag{12}$$

Six typical metrics are employed to assess the performance of the models, with higher values indicating better performance. Among these, four metrics are utilized for evaluating target change: Precision (Pre.), recall (Rec.), F1-score, and intersection over union (IoU). In addition, two metrics are employed to evaluate the overall classification accuracy: Overall accuracy (OA) and kappa coefficient. Formally, six evaluation metrics are defined as follows:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{13}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{14}$$

$$\mathrm{F1} = \frac{2}{\mathrm{Precision}^{-1}+\mathrm{Recall}^{-1}} \tag{15}$$

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{16}$$

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

$$\mathrm{Kappa} = \frac{\mathrm{OA}-\mathrm{CA}}{1-\mathrm{CA}} \tag{18}$$

where TP, TN, FP, and FN represent the quantities of true positives, true negatives, false positives, and false negatives, respectively. CA denotes the hypothetical probability of chance agreement between predictions and actual values, which can be

TABLE II
COMPARISONS ON MODEL ARCHITECTURE AND EFFICIENCY

Methods	Backbone	Multi-scale modeling	Same-scale interaction	Cross-scale interaction	FLOPs (G)	Params. (M)	Time (s)
FC-Siam-diff [27]	UNet	✓	✓	×	2.33	1.35	84
FC-Siam-conc [27]	UNet	✓	✓	×	2.33	1.55	72
STANet [47]	ResNet18	×	×	×	18.03	16.94	76
DTCDSCN [36]	SE-ResNet34	✓	✓	×	10.88	41.07	81
BIT [48]	ResNet18	×	×	×	25.92	11.99	110
DPCCNet [32]	ResNet50	✓	✓	×	33.68	13.74	292
SAGNet [38]	ResNet34	✓	✓	✓	12.25	32.23	125
BASNet(Ours)	ResNet18	✓	✓	✓	4.58	4.70	97

The used images are of size 3×256×256.

TABLE III
QUANTITATIVE COMPARISONS

Methods	LEVIR-CD	CDD	GZ-CD
	Pre. / Rec. / F1 / IoU / OA / Kappa	Pre. / Rec. / F1 / IoU / OA / Kappa	Pre. / Rec. / F1 / IoU / OA / Kappa
FC-Siam-diff	89.22 / 80.42 / 84.59 / 73.29 / 98.51 / 83.81	89.98 / 63.53 / 74.47 / 59.32 / 94.86 / 71.71	84.20 / 58.76 / 69.22 / 52.93 / 95.32 / 66.77
FC-Siam-conc	88.17 / 84.64 / 86.37 / 76.01 / 98.64 / 85.65	89.74 / 60.49 / 72.26 / 56.57 / 94.52 / 69.35	88.34 / 61.40 / 72.45 / 56.80 / 95.82 / 70.26
STANet	90.53 / 84.68 / 87.51 / 77.79 / 98.77 / 86.86	96.92 / 90.65 / 93.68 / 88.10 / 98.56 / 92.86	**93.33** / 69.18 / 79.46 / 65.92 / 96.80 / 77.76
DTCDSCN	90.70 / 87.17 / 88.90 / 80.02 / 98.89 / 88.32	96.32 / 92.42 / 94.33 / 89.27 / 98.69 / 93.59	83.43 / 76.04 / 79.57 / 66.07 / 96.50 / 77.66
BIT	91.12 / 87.61 / 89.33 / 80.72 / 98.93 / 88.77	**97.25** / 94.28 / 95.74 / 83.74 / 98.03 / 90.05	89.00 / 80.63 / 84.61 / 73.32 / 97.37 / 83.18
DPCCNet	90.83 / **89.32** / 90.07 / 81.94 / 99.00 / 89.55	96.85 / 94.80 / 95.92 / 91.96 / 99.02 / 95.26	88.89 / 81.92 / 85.26 / 74.31 / 97.46 / 83.88
SAGNet	91.33 / 82.93 / 86.93 / 76.88 / 98.73 / 86.26	96.59 / 95.33 / 95.96 / 92.23 / 99.05 / 95.42	87.10 / 72.82 / 79.32 / 65.73 / 96.60 / 77.49
BASNet (ours)	**92.66** / 88.81 / **90.69** / **82.96** / **99.07** / **90.21**	96.76 / **96.30** / **96.53** / **93.29** / **99.18** / **96.07**	88.21 / **84.22** / **86.17** / **75.70** / **97.58** / **84.84**

The highest scores are marked in bold. All scores are described in percentage (%).

formulated as

$$\mathrm{CA} = \frac{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})+(\mathrm{FN}+\mathrm{TN})(\mathrm{TP}+\mathrm{TN})}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})^2}. \tag{19}$$

C. Comparative Experiments

BASNet is evaluated by qualitative and quantitative comparisons with seven competitive CD methods. There are two categories of comparison methods: the first category includes FC-Siam-diff, FC-Siam-conc [27], DTCDSCN [36], and SAGNet [38], which are based on CNN and traditional attention mechanisms; the second category includes STANet [47], BIT [48], and DPCCNet [32], which combine CNN and transformer mechanisms. The specific properties of different methods are shown in Table II. In terms of architecture, our BASNet has all the advantageous properties to ensure appealing performance, including overall multiscale modeling, same-scale and cross-scale interaction. In contrast, most competitors lack consideration for certain properties. In terms of efficiency, we compare the computational efficiency of our proposed model with other models based on parameters (Params.), floating-point operations (FLOPs), and average time to train a single epoch. It can be seen that BASNet has lower computational cost and is more efficient than some state-of-the-art methods.

1) On LEVIR-CD: In Table III, BASNet achieves the best performance in terms of precision, F1-score, IoU, OA, and kappa. While its recall is slightly inferior to DPCCNet. Both BIT and DPCCNet exhibit high levels of precision and recall, indicating that the SA mechanism is beneficial for enhancing the global perception, reducing false positives and false negatives. STANet lacks bitemporal interaction, resulting in poor anti-interference ability. DTCDSCN and SAGNet, lacking in global perception, tend to miss detection of large objects. FC-Siam-conc and FC-Siam-diff underperform across various metrics, suggesting that simple bitemporal interaction methods, such as channel concatenation and elementwise subtraction might inadvertently confuse bitemporal semantics when handling significantly different image pairs.

Visual comparisons of six typical cases are illustrated in Fig. 7. Among these, Fig. 7(a)–(b) was captured under different lighting conditions, where ground reflection and tree shadows near the buildings can interfere with detection. However, effective bitemporal interaction and global perspective of BASNet mitigate the negative impact of lighting changes and ground color. In Fig. 7(c)–(d), small objects such as small houses with colors similar to the land and expanded factories are present. Most compared methods struggle to capture these nuances, whereas the parallel multiscale element addition and subtraction of BASNet can enhance homogeneous features and differentiating features, respectively, enabling the detection of subtle areas of change. In Fig. 7(e)–(f), where the changed areas are substantial, the compared methods demonstrate weak performance in handling intraclass inconsistencies, resulting in a significant number of false positives at the edges. In contrast, BASNet through the fusion and mutual guidance of multiscale features, enhances the correlation between pixels, mitigating most of the intraclass inconsistencies.

2) On CDD: In Table III, BASNet achieves the best performance in terms of recall, F1-score, IoU, OA, and kappa, while slightly trailing behind BIT in precision. Although BIT exhibits high precision, it is prone to the influence of dataset sample imbalance, leading to lower recall. The multiscale spatiotemporal attention of STANet effectively focuses on objects of various

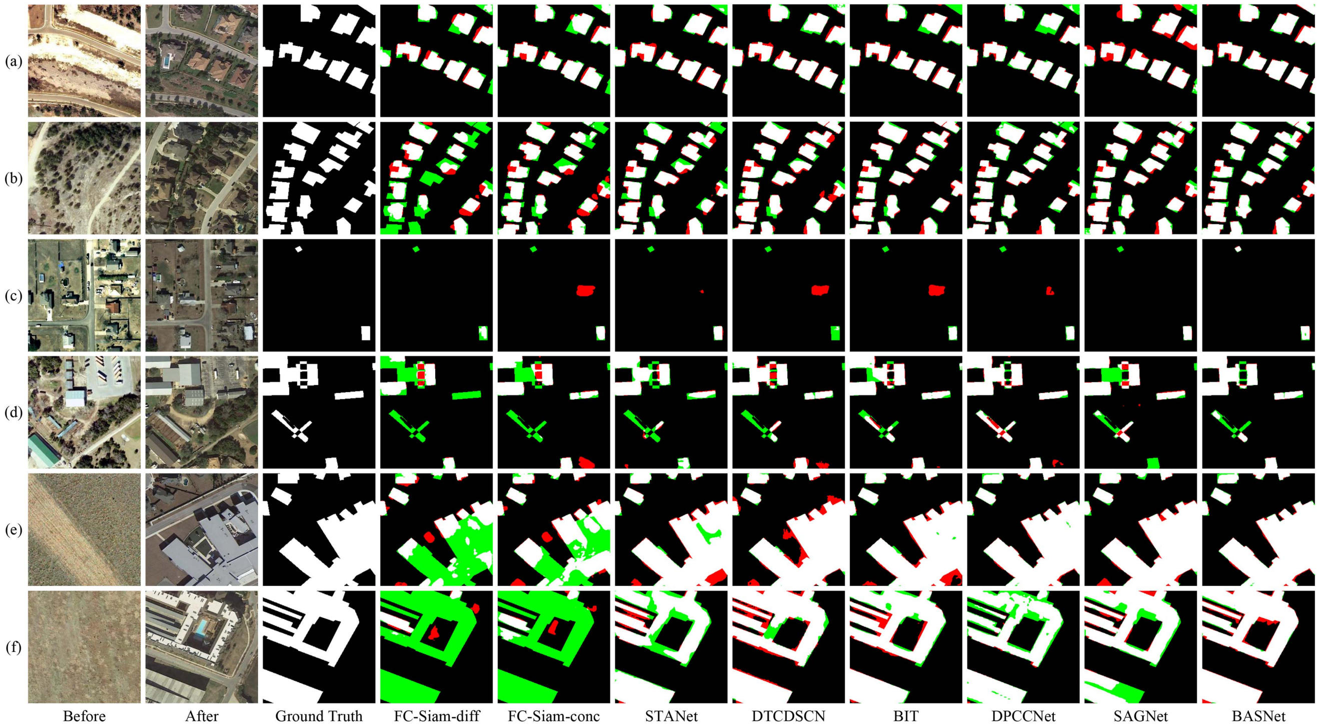

Fig. 7. Visualization results of different methods on LEVIR-CD. (a)–(f) denote prediction results of all the compared methods for different samples, respectively. In color classification, white for true positive, black for true negative, red for false positive, and green for false negative.

sizes, but its limited depth hampers performance. DTCDSCN and SAGNet demonstrate overall high accuracy, indicating that traditional channel attention and multiscale fusion are advantageous for segmenting diverse small targets. However, their global retrieval capability still lags behind DPCCNet and BASNet, which leverage bitemporal SA interaction. FC-Siam-conc and FC-Siam-diff suffer from severe intraclass inconsistencies and perform poorly, especially on CDD with higher target diversity.

Visual comparisons of six typical cases are illustrated in Fig. 8. In Fig. 8(a)–(b), where the primary targets are buildings with larger dimensions, BASNet effectively avoids interference from shadows of houses and trees during edge detection, while fully mitigating intraclass inconsistencies when handling extensive changes. In Fig. 8(c)–(d), focused on vehicles with smaller sizes, BASNet is capable of delineating the outlines of differently colored cars clearly. In Fig. 8(e)–(f), encompassing two different types of roads, snowy and dirt roads, BASNet showcases robust pixel-level recognition ability, allowing it to identify the distinct reflection and shadows caused by snowy roads. Its powerful generalization capability enables it to perform highly accurately in dealing with various road types.

3) On GZ-CD: In Table III, BASNet achieves the best performance in terms of recall, F1-score, IoU, OA, and kappa, while its precision performance is moderate. A cross-sectional comparison of results on the large-scale datasets reveals that STANet is significantly affected by sample imbalance, leading to a severe P-R imbalance, and the F1-score is greatly negatively impacted. Due to the limited labeled samples, the accuracy of BIT significantly decreases, and the predictive performance is far less effective than the other two datasets. This phenomenon indicates that, compared to purely CNN-based methods, transformer-based methods rely more on a large amount of labeled data and pretrained weights.

Visual comparisons of six typical cases are illustrated in Fig. 9. In Fig. 9(a)–(b), significant seasonal and lighting differences are evident. In Fig. 9(c)–(d), the detected objects are small, with similar colors causing interference. In Fig. 9(e)–(f), within a large construction area, structural differences and color variations result in intraclass inconsistency, while edge shadows lead to interclass ambiguity. Through comparison, BASNet can quickly learn and adapt to significant features of target changes from a small amount of labeled data, demonstrating stronger generalization than typical transformer-based models. This is because BASNet does not rely on the transformer for encoding and decoding such as BIT but only uses it as a medium for bitemporal attention interaction.

D. Ablation Study

To validate the effectiveness of BASNet, we conduct ablation experiments by adding or removing the proposed modules while keeping the backbone network unchanged. The experimental results are presented in Table IV. In addition, Fig. 10 visually demonstrates the impact of the proposed modules on feature mappings using channel visualization. Different scale features are upsampled to their original image size.

1) Effect of BAS Module: Compared with the network (a), the F1 and IoU of (b) are increased by 0.92% and 1.49%, respectively, indicating that adding this module on the basis

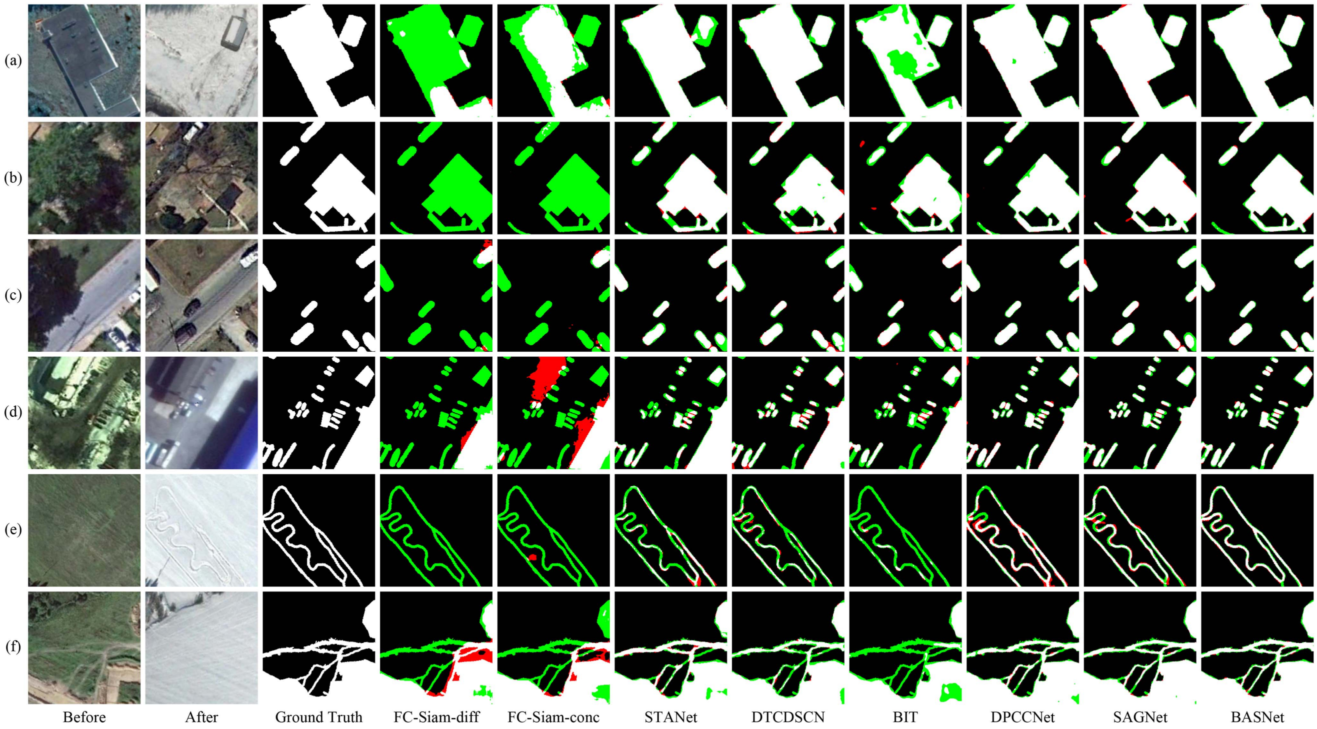

Fig. 8. Visualization results of different methods on CDD. (a)–(f) denote prediction results of all the compared methods for different samples, respectively. In color classification, white for true positive, black for true negative, red for false positive, and green for false negative.

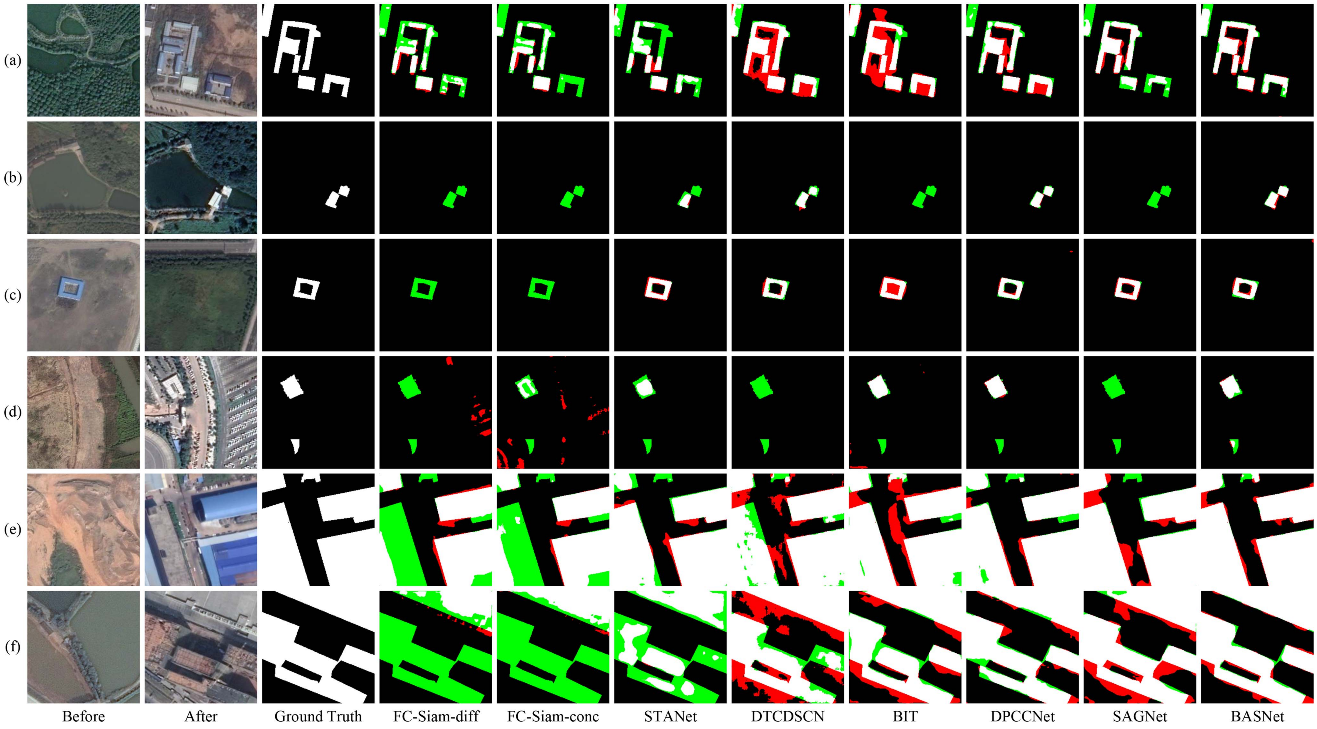

Fig. 9. Visualization results of different methods on GZ-CD. (a)–(f) denote prediction results of all the compared methods for different samples, respectively. In color classification, white for true positive, black for true negative, red for false positive, and green for false negative.

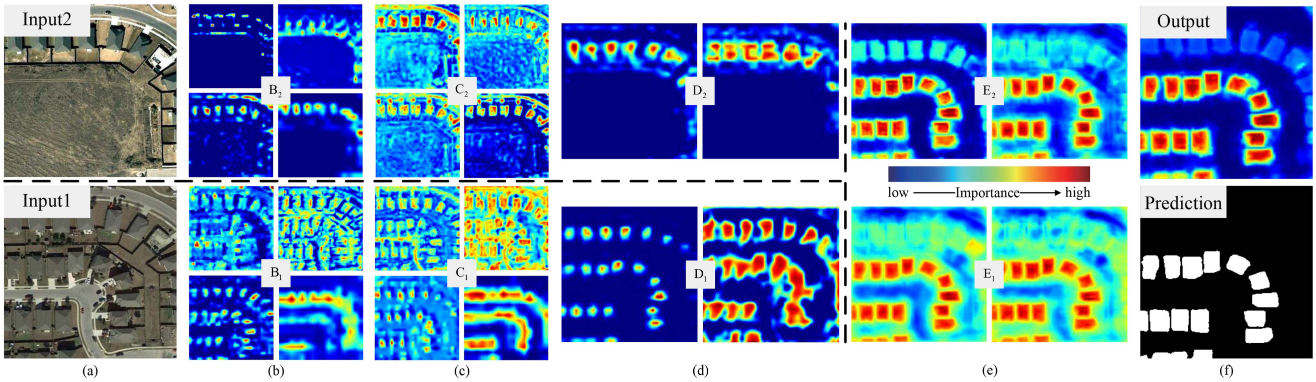

Fig. 10. Example of network visualization. (a) Input image. (b) Selected multiscale bitemporal features generated by the backbone network. (c) Selected interaction features produced by the BAS module. (d) Selected bitemporal features generated by the CAG module. (e) Selected interaction features produced by the WF module. (f) Change probability map and prediction map of the final output.

TABLE IV
ABLATION STUDY OF THE PROPOSED MODULES ON LEVIR-CD

Network	Modules			Results	
	BAS	CAG	WF	F1 (%)	IoU (%)
(a)	×	×	×	87.93	78.45
(b)	✓	×	×	88.85	79.94
(c)	×	✓	×	88.29	79.02
(d)	×	×	✓	88.58	79.53
(e)	×	✓	✓	89.98	81.79
(f)	✓	×	✓	90.42	82.52
(g)	✓	✓	×	90.14	82.05
(h)	✓	✓	✓	90.69	82.96

TABLE V
ABLATION STUDY ON SAME-SCALE ATTENTION SHARING MODULES

Modules	LEVIR-CD	CDD	GZ-CD
None	89.98	94.99	85.01
SA [39]	90.35	95.50	85.39
CA [56]	90.48	95.72	85.58
BAS (ours)	**90.69**	**96.53**	**86.17**

The bold entities indicate the optimal.

of the backbone network can effectively link the bitemporal information and improve the network performance. Compared with the network (h), the F1 and IoU of (e) are reduced by 0.71% and 1.17%, respectively, indicating that the BASNet network relies heavily on the BAS mechanism. For visualization, in Fig. 10(b), B_1 and B_2 represent bitemporal multiscale features encoded by the ResNet18 backbone. The attention distribution of features appears relatively disordered. After the initial encoding, the features undergo BAS, resulting in C_1 and C_2 in Fig. 10(c). This emphasizes buildings to a greater extent, and the shared bitemporal attention leads to similar attention distributions between the two. Further, Table V compares multiple same-scale attention methods on multiple datasets. Among them, SA [39] aims to use the SA mechanism independently in each tense, while cross-attention (CA) [56] is to exchange bitemporal query vectors on the basis of SA. Thanks to the anti-interference brought by the attention sharing mechanism, the performance of BAS in training small-scale data and complex target data is significantly improved compared with CA and SA.

TABLE VI
ABLATION STUDY ON CROSS-SCALE ATTENTION SHARING MODULES

Modules	LEVIR-CD	CDD	GZ-CD
None	90.42	95.60	84.87
CAB [57]	90.58	96.06	84.95
BGA [58]	90.61	96.44	85.53
CAG (ours)	**90.69**	**96.53**	**86.17**

The bold entities indicate the optimal.

2) *Effect of CAG Module:* Compared with the network (a), the F1 and IoU of (c) are increased by 0.36% and 0.57%, respectively, indicating that this module can help low-level features to obtain the target attention ability of high-level features. Compared with the network (h), the F1 and IoU of (f) are reduced by 0.27% and 0.44%, respectively, indicating that CAG is very important to improve the utilization efficiency of hierarchical features. For visualization, low-level features are guided by the attention of adjacent high-level features, producing features D_1 and D_2 in Fig. 10(d). Clearly, background changes are suppressed, but the attended areas do not fully cover the targets.Further, Table VI compares multiple cross-scale attention sharing modules on multiple datasets, where channel attention block (CAB) [57] aims to change the stagewise functional weights to enhance consistency, and bilateral guided aggregation (BGA) [58] combines feature representations at different scales by compensating for semantic and resolution gaps. Compared with them, CAG has no obvious advantage in improving accuracy, but due to the use of deep separable convolution instead of general convolution, the computational cost is greatly reduced when multiplexed multiple times.

3) *Effect of WF Module:* Compared with the network (a), the F1 and IoU of (d) are increased by 0.65% and 1.08%, respectively, indicating that the feature fusion mechanism of this module can improve the accuracy of feature recovery during upsampling. Compared with the network (h), the F1 and IoU of (g) are reduced by 0.55% and 0.91%, respectively, indicating that the feature fusion decoding of WF is an indispensable

part of the network codec structure. For visualization, multiscale WF can identify the semantic information of similar details and different details of multiscale features, and locate the contour of the target completely, so as to obtain E_1 and E_2 in Fig. 10(e).

V. Discussion

Although our proposed supervised model has achieved remarkable results in CD tasks, we must recognize that it still relies on a large amount of labeled data. There is still room for improvement in big data acquisition and model training methods. In response to this problem, we consider the advantages of unsupervised learning methods. Unsupervised CD methods usually rely on the statistical characteristics and spatiotemporal information of remote sensing images without premarked change information. Among them, pixel-based methods, such as threshold-based methods and clustering methods, are the most common. Threshold-based methods usually distinguish between changed and unchanged pixels by setting thresholds, but they may be sensitive to factors, such as illumination and noise. The clustering method attempts to cluster the image pixels into groups with similar features, and use the clustering results to detect changes. However, these methods often fail to make full use of the spatial relationship between pixels.

In view of the previous, we believe that it is necessary to explore how to combine supervised and unsupervised learning methods in future work. By introducing semisupervised learning, transfer learning, or weakly supervised learning, we can reduce the dependence on massive labeled data and improve model performance and generalization ability. At the same time, we can also consider introducing domain adaptation and data enhancement techniques to further enhance the generalization ability of the model, so that it can show better robustness and effect in different scenarios and data distributions. These efforts will help our model achieve better performance in future applications.

VI. Conclusion

In this article, we address the challenges of high computational cost and lack of effective bitemporal interaction in existing CD methods by proposing an efficient model, BASNet. It is composed of a backbone network, BAS module for bitemporal interaction, CAG module for same-scale guidance, and WF module for cross-scale integration. Specifically, we first encode the bitemporal remote sensing images into coarse-grained features using the backbone network, generating multiscale representations. BAS then optimizes the arrangement of bitemporal SA through the construction of a shared global attention. Within CAG, high-level features enhance low-level features by leveraging their attention weights, enhancing the efficiency of multiscale utilization. Next, bitemporal feature interaction is achieved through elementwise addition and subtraction. Finally, WF effectively strengthens detection capabilities for small objects and edge information through multiscale fusion. Experimental results demonstrate that BASNet achieves excellent performance on three public datasets with low computational cost.

References

[1] 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.

[2] L. Weng, K. Pang, M. Xia, H. Lin, M. Qian, and C. Zhu, “Sgformer: A local and global features coupling network for semantic segmentation of land cover,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 6812–6824, 2023.

[3] H. Ji, M. Xia, D. Zhang, and H. Lin, “Multi-supervised feature fusion attention network for clouds and shadows detection,” *ISPRS Int. J. Geo-Inf.*, vol. 12, p. 247, 2023.

[4] Z. Hu, L. Weng, M. Xia, K. Hu, and H. Lin, “HyCloudX: A multi-branch hybrid segmentation network with band fusion for cloud/shadow,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 6762–6778, 2024.

[5] J. Chen, M. Xia, D. Wang, and H. Lin, “Double branch parallel network for segmentation of buildings and waters in remote sensing images,” *Remote Sens.*, vol. 15, no. 6, p. 1536, 2023.

[6] H. Qiao, X. Wan, Y. Wan, S. Li, and W. Zhang, “A novel change detection method for natural disaster detection and segmentation from video sequence,” *Sensors*, vol. 20, no. 18, p. 5076, 2020.

[7] X. Li, F. Ling, G. M. Foody, and Y. Du, “A superresolution land-cover change detection method using remotely sensed images with different spatial resolutions,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 7, pp. 3822–3841, Jul. 2016.

[8] C. Marin, F. Bovolo, and L. Bruzzone, “Building change detection in multitemporal very high resolution SAR images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2664–2682, May 2015.

[9] L. J. Jansen and A. Di Gregorio, “Parametric land cover and land-use classifications as tools for environmental change detection,” *Agriculture, Ecosystems Environ.*, vol. 91, no. 1–3, pp. 89–100, 2002.

[10] A. Maćkiewicz and W. Ratajczak, “Principal components analysis (PCA),” *Comput. Geosciences*, vol. 19, no. 3, pp. 303–342, 1993.

[11] E. F. Lambin and A. H. Strahlers, “Change-vector analysis in multitemporal space: A tool to detect and categorize land-cover change processes using high temporal-resolution satellite data,” *Remote Sens. Environ.*, vol. 48, no. 2, pp. 231–244, 1994.

[12] Y. Tang, L. Zhang, and X. Huang, “Object-oriented change detection based on the Kolmogorov–Smirnov test using high-resolution multispectral imagery,” *Int. J. Remote Sens.*, vol. 32, no. 20, pp. 5719–5740, 2011.

[13] F. Bovolo, L. Bruzzone, and M. Marconcini, “A novel approach to unsupervised change detection based on a semisupervised SVM and a similarity measure,” *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 7, pp. 2070–2082, Jul. 2008.

[14] J. Im and J. R. Jensen, “A change detection model based on neighborhood correlation image analysis and decision tree classification,” *Remote Sens. Environ.*, vol. 99, no. 3, pp. 326–340, 2005.

[15] M. Hussain, D. Chen, A. Cheng, H. Wei, and D. Stanley, “Change detection from remotely sensed images: From pixel-based to object-based approaches,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 80, pp. 91–106, 2013.

[16] P. Xiao, X. Zhang, D. Wang, M. Yuan, X. Feng, and M. Kelly, “Change detection of built-up land: A framework of combining pixel-based detection and object-based recognition,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 119, pp. 402–414, 2016.

[17] X. Zhang, S. Cheng, L. Wang, and H. Li, “Asymmetric cross-attention hierarchical network based on CNN and transformer for bitemporal remote sensing images change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 2000415.

[18] W. Ren, Z. Wang, M. Xia, and H. Lin, “MFINet: Multi-scale feature interaction network for change detection of high-resolution remote sensing images,” *Remote Sens.*, vol. 16, no. 7, 2024.

[19] M. Lin, G. Yang, and H. Zhang, “Transition is a process: Pair-to-video change detection networks for very high resolution remote sensing images,” *IEEE Trans. Image Process.*, vol. 32, pp. 57–71, 2023.

[20] Z. Lv, H. Huang, L. Gao, J. A. Benediktsson, M. Zhao, and C. Shi, “Simple multiscale unet for change detection with heterogeneous remote sensing images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 2504905.

[21] X. Wang et al., “A high-resolution feature difference attention network for the application of building change detection,” *Int. J. Appl. Earth Observ. Geoinformation*, vol. 12, Aug. 2022, Art. no. 102950.

[22] D. Wang, L. Weng, M. Xia, and H. Lin, “MBCNet: Multi-branch collaborative change-detection network based on siamese structure,” *Remote Sens.*, vol. 15, no. 9, p. 2237, 2023.

[23] B. Chen, M. Xia, M. Qian, and J. Huang, "MANet: A multi-level aggregation network for semantic segmentation of high-resolution remote sensing images," *Int. J. Remote Sens.*, vol. 43, no. 15–16, pp. 5874–5894, 2022.

[24] Z. Wang, M. Xia, L. Weng, K. Hu, and H. Lin, "Dual encoder–decoder network for land cover segmentation of remote sensing image," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 2372–2385, 2024.

[25] Z. Song, X. Li, R. Zhu, Z. Wang, Y. Yang, and X. Zhang, "ERMF: Edge refinement multi-feature for change detection in bitemporal remote sensing images," *Signal Process. Image Commun.*, vol. 116, 2023, Art. no. 116964.

[26] C. Ma, L. Weng, M. Xia, H. Lin, M. Qian, and Y. Zhang, "Dual-branch network for change detection of remote sensing image," *Eng. Appl. Artif. Intell.*, vol. 123, 2023, Art. no. 106324.

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

[28] L. Song, M. Xia, J. Jin, M. Qian, and Y. Zhang, "SUACDNet: Attentional change detection network based on siamese U-shaped structure," *Int. J. Appl. Earth Observation Geoinformation*, vol. 105, Dec. 2021, Art. no. 102597.

[29] 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, Jan. 2021.

[30] 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, Jan. 2022, Art. no. 8007805.

[31] 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, vol. 37, no. 12, pp. 14187–14195.

[32] Q. Shu, J. Pan, Z. Zhang, and M. Wang, "DPCC-Net: Dual-perspective change contextual network for change detection in high-resolution remote sensing images," *Int. J. Appl. Earth Observation Geoinformation*, vol. 112, 2022, Art. no. 102940.

[33] H. Ren, M. Xia, L. Weng, K. Hu, and H. Lin, "Dual-attention-guided multiscale feature aggregation network for remote sensing image change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 4899–4916, 2024.

[34] A. Varghese, J. Gubbi, A. Ramaswamy, and P. Balamuralidhar, "Changenet: A deep learning architecture for visual change detection," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2018.

[35] 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, Aug. 2020.

[36] 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.

[37] S. Fang, K. Li, and Z. Li, "Changer: Feature interaction is what you need for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, May 2023, Art. no. 5610111.

[38] H. Yin et al., "Attention-guided siamese networks for change detection in high resolution remote sensing images," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 117, 2023, Art. no. 103206.

[39] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6000–6010.

[40] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2021, *arXiv:2010.11929*.

[41] K. Chen, X. Dai, M. Xia, L. Weng, K. Hu, and H. Lin, "Msfanet: Multi-scale strip feature attention network for cloud and cloud shadow segmentation," *Remote Sens.*, vol. 15, p. 4853, 2023.

[42] X. Dai, K. Chen, M. Xia, L. Weng, and H. Lin, "LPMSNet: Location pooling multi-scale network for cloud and cloud shadow segmentation," *Remote Sens.*, vol. 15, p. 4005, 2023.

[43] L. Ding, M. Xia, H. Lin, and K. Hu, "Multi-level attention interactive network for cloud and snow detection segmentation," *Remote Sens.*, vol. 16, no. 1, p. 112, 2024.

[44] 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.

[45] 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, 2022, Art. no. 5224713.

[46] C. Zhang, L. Weng, L. Ding, M. Xia, and H. Lin, "CRSNet: Cloud and cloud shadow refinement segmentation networks for remote sensing imagery," *Remote Sens.*, vol. 15, no. 6, 2023.

[47] 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, p. 1662, 2020.

[48] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5607514.

[49] L. Song, M. Xia, L. Weng, H. Lin, M. Qian, and B. Chen, "Axial cross attention meets CNN: Bibranch fusion network for change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 21–32, Jan. 2023.

[50] Y. Feng, J. Jiang, H. Xu, and J. Zheng, "Change detection on remote sensing images using dual-branch multilevel intertemporal network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4401015.

[51] Z. Ma, M. Xia, H. Lin, M. Qian, and Y. Zhang, "FENet: Feature enhancement network for land cover classification," *Int. J. Remote Sens.*, vol. 44, no. 5, pp. 1702–1725, 2023.

[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, p. 1662, May 2020.

[53] 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.

[54] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, H. Ding, and X. Huang, "SemiCDNet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5891–5906, Jul. 2021.

[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017, *arXiv:1412.6980*.

[56] C.-F. R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 347–356.

[57] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *Proc. lEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018.

[58] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, pp. 3051–3068, 2021.

Zhongchen Wang received the B.S. degree in automation from the Nanjing Institute of Technology, Nanjing, China, in 2022, and the master's degree in electronic information from the Nanjing University of Information Science and Technology, Nanjing, China.

His research interests include deep learning and its applications.

Guowei Gu received the B.S. degree from the Huaiyin Institute of Technology, Huaian, China, in 2022, and the master's degree in electronic information from the Nanjing University of Information Science and Technology, Nanjing, China.

His research interests include deep learning and its applications.

Min Xia (Member, IEEE) received the Ph.D. degree in cybernetics control engineering from Donghua University, Shanghai, China, in 2009.

He is currently a Professor with the Nanjing University of Information Science and Technology, Nanjing, China. He is also the Deputy Director of Jiangsu Key Laboratory of Big Data Analysis Technology, Nanjing, China. His research interests include machine learning theory and its application.

Liguo Weng received the Ph.D. degree in electrical engineering from North Carolina A&T State University, Greensboro, NC, USA, in 2010.

He is currently an Associate Professor with the College of Automation, Nanjing University of Information Science and Technology, Nanjing, China. His research interests include deep learning and its application in remote sensing image analysis.

Kai Hu received the bachelor's degree from the China University of Metrology, Hangzhou, China, in 2003, the master's degree from the Nanjing University of Information Science and Technology, Nanjing, China, in 2008, and the Ph.D. degree in instrument science and engineering from Southeast University, Nanjing, China, in 2015.

He is currently an Associate Professor with the Nanjing University of Information Science and Technology. His research interests include deep learning and its applications in remote sensing images.