Exploring Context Alignment and Structure Perception for Building Change Detection

Qi Wang, *Senior Member, IEEE*, Mingwei Zhang, Jiawei Ren, Qiang Li, *Member, IEEE*

***Abstract*—Automatically monitoring building changes can assist human experts in disaster rescue, urban planning, and resource protection. Consequently, much research focus on building change detection. Recently, the methods based on deep learning have achieved impressive performance. However, most of them ignore the effect of bitemporal image misalignment, which is prone to lead to false detection. To this end, a building change detection model with Context Alignment and Structure Perception (CASP) is proposed. First, imitating the brain logic of humans to identify changes, a bitemporal interactive alignment module is designed, which suppresses the spatial dislocation noise via a bidirectional reference-guided feature aggregation strategy. Building on this, a difference-induced alignment module is introduced to mitigate the adverse impact of misalignment errors further and improve the accuracy of building change detection. Second, a structure-aware feature fusion module is developed and integrated into the feature encoder, to enhance the discrimination of building representations and highlight the specificity of the proposed method. Extensive experiments on three representative building change detection datasets are implemented to verify the superiority of the above improvements. The quantitative and qualitative results demonstrate the proposed method achieves competitive performance. The code is available at https://github.com/ptdoge/CASP.**

***Index Terms*—Building change detection, context alignment, structure perception, feature refinement.**

I. Introduction

BUILDING change detection is a crucial and practical task in computer vision and remote sensing. It aims to monitor the demolition, construction, and updating of buildings at a region of concern via remote sensing images captured at different times. It can be applied to illegal construction surveys [1], disaster damage detection [2], and ecological resource protection [3]. Looking into the changed buildings from the large-scale remote sensing images is labor-intensive and time-consuming. Therefore, achieving automatic change detection of buildings has been a hot research topic for a long time. In particular, recent methods based on data-driven deep learning [4] have achieved impressive results.

In the early days, data is expensive and the computational power is limited. Most researchers focus on exploring unsupervised change detection methods from pixel to object levels by employing machine learning techniques. Many classical models such as change vector analysis [5], post-classification comparison [6], and principal component analysis-based change detection [7] are proposed. However, these methods fail to filter building-irrelevant changes. Meanwhile, they are fragile to the variation of imaging conditions, which can be attributed to the high sensitivity of manual-designed features in them. In the past decade, with the development of hardware resources and the accumulation of multi-temporal remote sensing data, change detection models based on deep learning have increasingly dominated the domain. These models can effectively alleviate the interference of pseudo-changes. Among them, some are tailored for building change detection. The rest can be easily adapted to detect changes in buildings by utilizing annotated building change detection datasets.

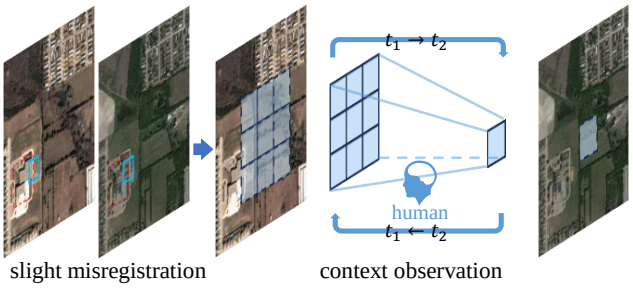

Fig. 1. Occasionally, there is a slight offset between the bitemporal images due to the incomplete accuracy of the registration algorithm. Most existing methods struggle with this challenge. In contrast, humans can precisely identify the real changes and filter misregistration errors in bitemporal images by bidirectionally observing the context similarity. In this work, a mechanism imitating the pattern of humans to alleviate the above problem is explored.

Most change detection models based on deep learning can be disassembled into three parts: bitemporal feature encoding, difference representation extraction, and change classification. Many convolutional neural networks (CNNs) or Transformers, such as ResNet [8], MobileNet [9], and Segformer [10], are off the shell. They are commonly used to encode bitemporal features from bitemporal images [11]–[13]. Researchers pay more attention to the modeling and decoding of change embedding. In detail, to model discriminative change representation and suppress pseudo-change noises, diverse bitemporal interaction mechanisms are investigated extensively [14]–[19]. For example, Zhang *et al.* [15] propose a difference-guided bitemporal aggregation strategy with the interaction between difference and bitemporal representations. Liu [20] *et al.* exploit the interaction of statistical knowledge in bi-temporal features to reduce spectral differences in unchanged regions. Changer proposed by Li *et al.* [16] explores an interaction mechanism

This work was supported by the National Natural Science Foundation of China under Grant 62301385, 62471394, and U21B2041.

Mingwei Zhang is with the School of Computer Science and the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P.R. China (e-mail: dlaizmw@gmail.com).

Qi Wang, Jiawei Ren, and Qiang Li are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P.R. China (e-mail: crabwq@gmail.com, jiaweiren2001@mail.nwpu.edu.cn, liqmges@gmail.com).

Qiang Li is the corresponding author.

with the exchange of the bitemporal features across both channel and spatial dimensions. Meanwhile, to effectively leverage contextual information for accurate change detection, various multi-scale feature decoding strategies are explored [21]–[24]. The decoder architectures like that in UNet [25] or FPN [26] are common. As for the change classification, most methods are classification-based, and a few are metric-based [15], [27]. However, these methods are dependent on well-aligned bitemporal images. They ignore the slight errors in some local regions claimed aligned, which is adverse to accurate change identification. To alleviate the above problem, how to further align the spatial contents at the feature level is explored in this work.

Humans precisely identify the changed regions and effectively exclude the pseudo-changed caused by unaligned errors by flexibly observing the contextual similarity and comparing the differences, as shown in Fig. 1. Inspired by human cognition, we model the observation logic of humans as a bidirectional reference-guided contextual aggregation. Thereby, a bitemporal interactive alignment module (BIAM) and a difference-induced alignment module (DIAM) are developed. They are used to align the multi-level features for reliable difference extraction. Besides, motivated by previous building change detection models [28]–[30], a structure-aware fusion module (SAFM) is designed. Unlike supervised edge extraction, it embeds implicit edge structure representation into the semantic features, thereby enhancing the discrimination of building representation. The proposed context alignment and structure perception modules are integrated into the pipeline of our model, as shown in Fig. 2(a), which includes four parts: feature encoder, context alignment, difference extraction, and change classifier. Thus, a novel model called CASP is proposed for building change detection. Overall, the contributions made by this article are as follows:

- We explore the context alignment strategy by analyzing the logic of humans for identifying changes. In detail, a bitemporal interactive alignment module and a difference-induced alignment module are developed to alleviate the negative effect of the slight spatial offset in bi-temporal images.
- We investigate the benefit of structure perception to enhance the discrimination of building representation. A structure-aware fusion module is designed to extract the edge structure representation and generate enhanced features, thus boosting the accuracy of building change detection.
- Experiments on three recognized building change detection datasets show that the proposed method is competitive. In particular, it has a clear advantage on the LEVIR-CD+ dataset, in which there are some typical misaligned examples.

II. Related Work

In this section, the previous works exploring the alignment methods for change detection at the image level or the feature level are introduced firstly. Then, the edge-assisted building change detection models are discussed.

A. Image and Feature Alignment for Change Detection

Spatial registration is an important prerequisite to achieving accurate bitemporal image change detection [31]. Most existing works assume that the bi-temporal images are strictly aligned by registration, ignoring possible errors. Recently, some studies attempt to alleviate the above problem from image and feature levels [32]–[34]. For example, to alleviate the geometric distortion caused by geolocation coordinate shifting and simplify the preprocessing pipeline of change detection, Zhao *et al.* [35] propose a self-supervised framework that can be reused for the bitemporal image alignment and the change detection tasks. Park *et al.* proposed SimSaC [32], a multi-task learning model, which can implement scene flow estimation and simultaneously detect changes. SimSac is designed to adapt to the numerous challenging scenes that are difficult to accurately register in real-world settings. These methods primarily aim to address significant deformation or shifts between bitemporal images. Additionally, some models investigate mechanisms to mitigate small offsets by refining bitemporal features. Tain *et al.* [36] propose a resolution- and alignment-aware network, where a deformable convolution unit [37], [38] is designed for bitemporal feature alignment. BiFA proposed by Zhang *et al.* [34] attempts to realize the spatial alignment between bitemporal images by estimating their offset at the feature level, mitigating the inadequate registration. Like the above works, Liu *et al.* [39] alleviates the effect of misregistration by introducing a multi-scale offset calibration module. This work also explores the bitemproal context alignment at the feature level, aiming at the inadequately accurate registration. In this work, we explore an interactive alignment strategy obeying human intuition. Furthermore, a progressive difference-induced alignment mechanism is developed by incorporating deformable convolution.

B. Edge-assisted Building Change Detection

Many works have explored how to utilize changed building edges to improve the accuracy of building change detection, which can be classified into two categories: edge prior introduction and edge-relevant feature enhancement [40]–[44]. About the former, Liang *et al.* [40] adopt a multi-task learning strategy to improve performance by simultaneous edge detection and building change detection. Bai *et al.* [45] employ a similar idea to improve the representation ability of changed buildings. Meanwhile, Chen *et al.* [46] propose EGDE-Net, a network designed to enhance both the integrity of change regions and the accuracy of boundaries. In this approach, prior information regarding boundaries, derived from the change ground truth, is utilized to supervise network training. The same strategy is employed in EGHNet [41], IBMNet [47], etc. Besides, without direct edge extraction, Eftekhari *et al.* [48] propose an edge-based consistency loss to converge the edges of the training and the testing samples. Overall, these works mainly introduce extra edge supervision to utilize edge prior, thereby improving the accuracy of building change detection. As for the latter, an edge-aware module and an edge-guided feature module are developed to reduce the blurred edges in EGPNet [43]. Meanwhile, an edge-aware module is designed for feature refinement in EGDE-Net [46]. Similarly, there is a refining block in EARTDer [44] to enhance the combined features of edge and change representation. These works aim

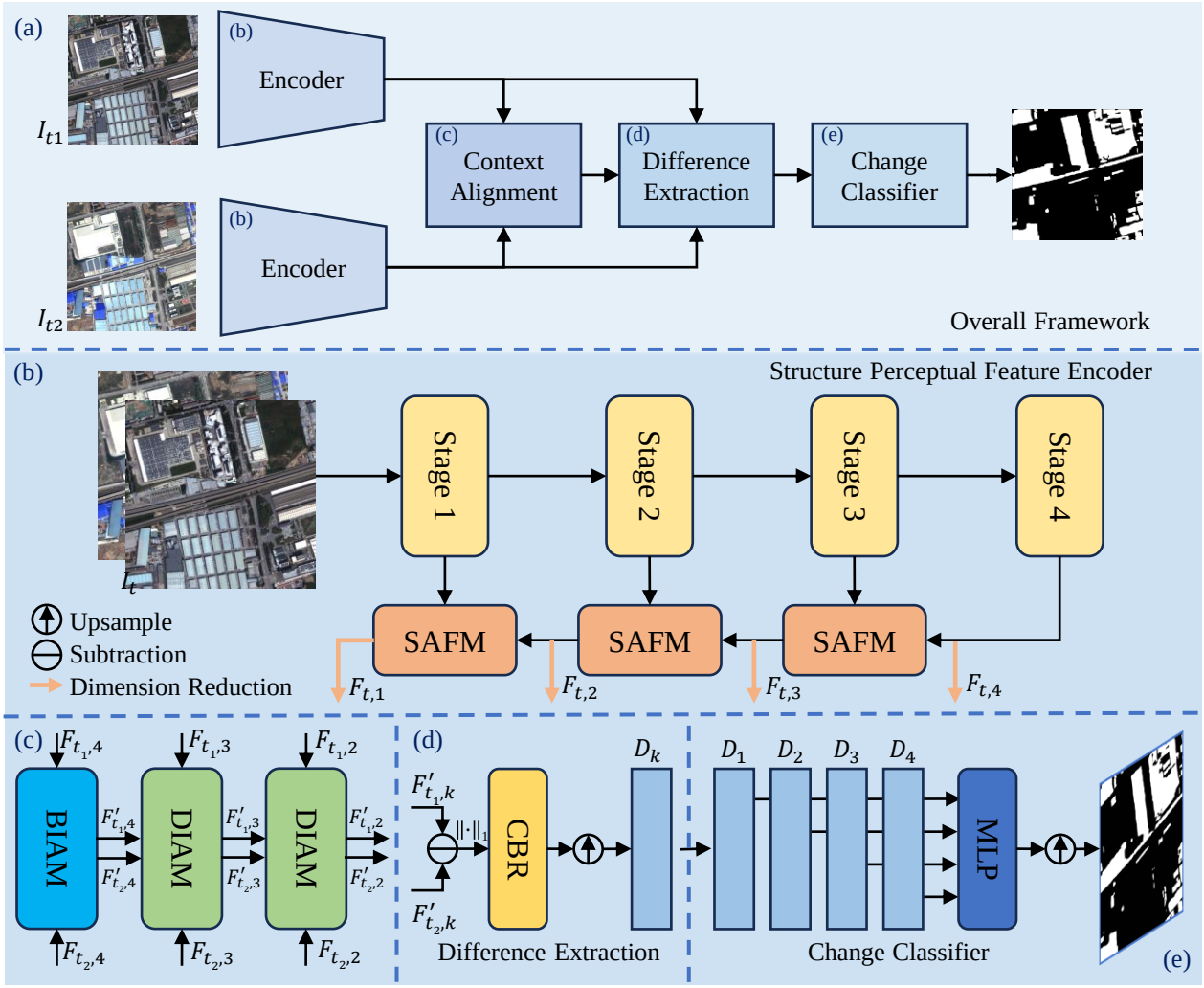

Fig. 2. The illustration of the proposed method. (a) The overall pipeline with bitemporal context alignment for building change detection. (b) Structure perceptual feature encoder. (c) Context alignment sub-network. (d) Difference extraction sub-network. (e) Change classifier.

to refine change representation via edge feature injection. According to the above studies about edge prior introduction and edge-guided feature refinement, it can be concluded that effectively mining the edge clues is helpful for building change detection. Therefore, an object structure perception technique is introduced in this work to explore the edge representation and improve the accuracy of building change detection.

III. Methodology

In this section, we first introduce the overall framework of the proposed method in detail from five aspects: feature encoder, context alignment, difference extraction, change classification, and loss function. Then, the structure-aware fusion module and the feature context alignment module developed are described in turn.

A. Framework

The overall framework of the proposed method CASP is presented in Fig. 2(a). As with many previous methods, a Siamese encoder is employed to extract multi-level features from bitemporal inputs respectively. Differently, a structure perception strategy for better building-relevant representation extraction is introduced in the encoder network. After that, refined bitemporal features are generated by the attentive processing in a context alignment sub-network. Finally, a simple difference extraction module and a lightweight change classifier are combined to identify changes. The detailed architectures of the procedures mentioned above are illustrated in Fig. 2(b), (c), (d), and (e) respectively.

Structure Perceptual Feature Encoder: As shown in Fig. 2(b), the multi-level features are firstly extracted from a popular backbone network (ResNet [8], MiT [10], [49], etc.) with four stages. Let f_i denote the feature at the i-th stage, where its height and width are each $\frac{1}{2^{i+1}}$ of the original image dimensions. Then, the SAFM is introduced to progressively fuse the features from adjacent levels, thereby obtaining the structure-enhanced features $f_i^{'}, i \in \{1,2,3\}$:

$$f_i^{'} = \mathrm{SAFM}(f_i, f_{i+1}^{'}),\ i \in \{1,2,3\}, \tag{1}$$

where $f_4^{'} = f_4$. To reduce the computational complexity of subsequent operations, a dimension-reduction (DR) technique

like previous works [15] is employed to transform the number of channels in $f_i^{'}$:

$$F_{t,i} = \mathrm{DR}(f_{t,i}^{'}),\ i \in \{1,2,3,4\},\ t \in \{t_1, t_2\}, \tag{2}$$

where the temporal index t is added to facilitate the following descriptions. $F_{t,i}$ is the final output of the feature encoder.

Context Alignment: To alleviate the impact of slight misalignment in bitemporal images, a context alignment subnetwork consisting of one BIAM and two DIAMs is developed to refine the bitemporal features. Its forward process is illustrated in Fig. 2(c) and can be formulated as follows:

$$F_{t_1,4}^{'}, F_{t_2,4}^{'} = \mathrm{BIAM}(F_{t_1,4}, F_{t_2,4}), \tag{3}$$

$$F_{t_1,3}^{'}, F_{t_2,3}^{'} = \mathrm{DIAM}(F_{t_1,3}, F_{t_2,3}, F_{t_1,4}^{'}, F_{t_2,4}^{'}), \tag{4}$$

$$F_{t_1,2}^{'}, F_{t_2,2}^{'} = \mathrm{DIAM}(F_{t_1,2}, F_{t_2,2}, F_{t_1,3}^{'}, F_{t_2,3}^{'}). \tag{5}$$

Here, $F_{t,i}^{'}$ refers to the align-refined feature. Notably, this sub-network is only used to process the middle and highest-level features for $i \in \{2,3,4\}$. Considering that the lowest-level feature $F_{t,1}$ primarily presents texture structure and lacks semantic information, context alignment is not applied to it. Otherwise, the structure-relevant features of changed buildings may be disrupted. For convenience in subsequent descriptions, let $F_{t,1}^{'} = F_{t,1}$.

Difference Extraction: The multi-level difference representations are acquired as follows:

$$D_i = \mathrm{CBR}(\|F_{t_1,i}^{'} - F_{t_2,i}^{'}\|_1)_{\uparrow 2^{i-1}}, \tag{6}$$

where D_i indicates the difference representation characterizing changes. CBR comprises one convolution operator with kernel size 3, one batch normalization layer, and the ReLU activation function.

Change Classifier: At last, the features concatenated at the channel dimension for change classification. Change classifier is a lightweight MLP like that in [10], [50]. Let P represent the predicted change map:

$$P = \mathrm{argmax}(\mathrm{MLP}(D_1 \| D_2 \| D_3 \| D_4)_{\uparrow 4}), \tag{7}$$

where $\|$ indicates the concatenation operation.

Loss Function: The overall pipeline of the proposed method is trained only by the cross-entropy loss:

$$\mathcal{L}_{ce} = -\sum_{i}^{N} y_i log(p_i) + (1-y_i) log(1-p_i), \tag{8}$$

where y_i denotes the ground truth label and p_i represents the change probability value of pixel i.

B. Structure-Aware Fusion Module

Building change detection is a typical single-class object change detection task. A key to improving the accuracy of identifying building changes lies in their discriminative modeling. Humans typically recognize a building by perceiving its structure-relevant features, such as contours, and inherent knowledge at the semantic level, distinguishing it from the surrounding background. Inspired by that, the SAFM is developed to inject structure-relevant information into multi-level features extracted from the backbone network. It is illustrated in Fig. 3. In detail, we employ a neighbor-level feature fusion strategy with differential convolution (DC) [51] to explore latent contour information of buildings. The DC aims to calculate the difference between features at the pixel (m,n) and its neighbor:

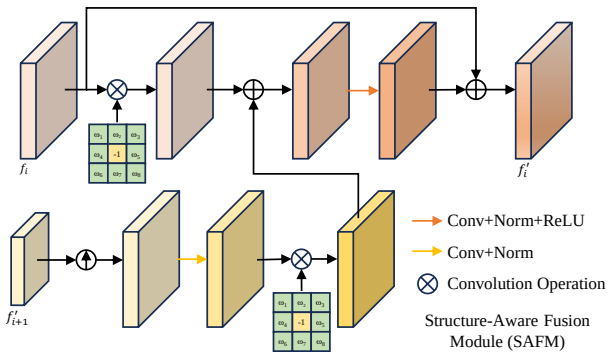

Fig. 3. The schematic diagram of the SAFM.

$$\hat{f} = \sum W(i,j)(f(i,j) - f(m,n)) \tag{9}$$

where $(i,j) \in \mathcal{N}_{(m,n)}$. $\mathcal{N}_{(m,n)}$ is the local neighbor region of pixel (m,n) and its size is 3×3. $W(i,j)$ is a learnable weight. Intuitively, it can reveal the boundary region at the semantic transition zone. Meanwhile, considering the focus ability of the deep layers on the interested object, $\hat{f}_{i+1}^{'}$ are extracted and fused with $\hat{f}_i$, and then they are injected into f_i to result in $f_i^{'}$. Accordingly, SAFM helps enhance the representation discrimination of buildings, thereby reducing the impact of other objects.

C. Feature Context Alignment

Due to the geolocation errors and the difference between imaging views in bitemporal images, registration is a necessary preprocessing for change detection. However, some slight misalignment is hard to eliminate owing to the inherent limitation of registration algorithms. Therefore, to improve the accuracy of building change detection, we investigate a registration noise suppression technique based on feature context alignment, which consists of the BIAM and the DIAM. The schematics of the BIAM and the DIAM are shown in Figs. 4 and 5 respectively.

Bitemporal Interactive Alignment Module: BIAM aims to model the observation logic of humans to discern misalignment and identify changes in bitemporal images. BIAM is a bidirectional feature interaction module between $F_{t_1,i}$ and $F_{t_2,i}$. For brevity, its operation is described below according to the direction modulating the $F_{t_1,i}$ to align the $F_{t_2,i}$. Firstly, the similarity between $F_{t_2,i}(m,n)$ at the position (m,n) and $F_{t_1,i}(k,l), (k,l) \in \mathcal{R}(m,n)$ is measured by a guided cross-attention mechanism:

$$c(k,l) = (WF_{t_1,i}(k,l))^T (UF_{t_2,i}(m,n)), \tag{10}$$

$$g(m,n) = (WF_{t_2,i}(m,n))^T (UF_{t_2,i}(m,n)), \tag{11}$$

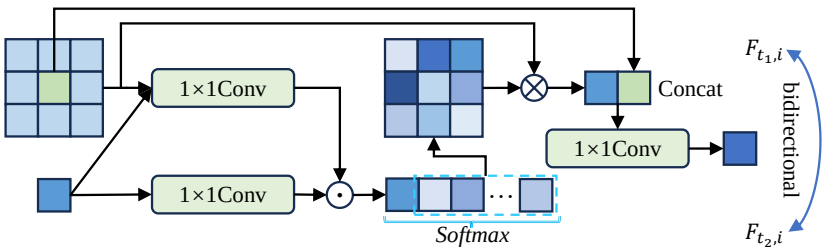

Fig. 4. The schematic diagram of the BIAM.

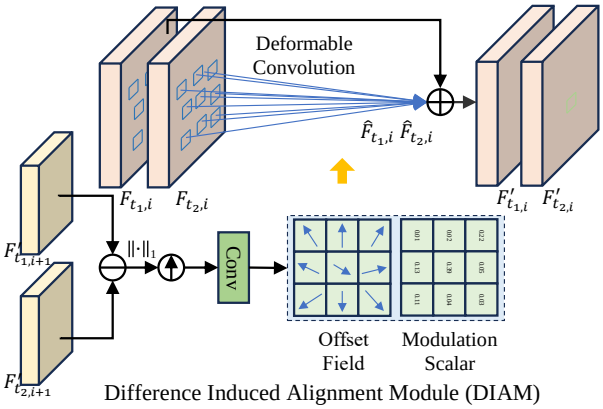

Fig. 5. The schematic diagram of the DIAM.

$$\omega = z(s([c(k,l),(k,l) \in \mathcal{R}(m,n)] \| g(m,n))), \tag{12}$$

where $c(k,l)$ is the cross-attention term and $g(m,n)$ is the guided term. $\mathcal{R}$ refers to the observed context scopes (OCS). W and U denote two linear transform layers. $s(\cdot)$ is the Softmax function, and $z(\cdot)$ means discarding the similarity value at the guided term. Notably, introducing $g(m,n)$ aims to suppress features whose semantics are inconsistent with $F_{t_2,i}(m,n)$. Afterward, $F_{t_1,i}$ in the context scopes observed at the position (m,n) is aggregated to get the aligned features $\hat{F}_{t_1,i}(m,n)$ with $F_{t_2,i}(m,n)$:

$$\hat{F}_{t_1,i}(m,n) = \sum \omega(k,l) F_{t_1,i}(k,l). \tag{13}$$

According to Eqs. 12 and 13, it can be inferred that the background interference will be amplified without the constraint of $g(m,n)$. At last, $\hat{F}_{t_1,i}$ are fused with its initial embedding $F_{t_1,i}$ to reduce the adverse effect of the registration noise and retain original detail information. The operation rectifying $F_{t_2,i}$ to align with $F_{t_1,i}$ is the same as the above description.

In summary, BIAM alleviates the content offset in bitemporal images by a locally guided cross-attention mechanism. Intuitively, the OCS at different levels should be distinct and increase with the feature size, To avoid introducing a greater computational burden, BIAM is only employed for the deepest feature layer ($i = 4$) in the designed model.

Difference-Induced Alignment Module: DIAM is developed to refine the features at other levels ($i \in \{2,3\}$) with lower computational complexity than BIAM. In detail, a difference-induced deformable convolution is designed in DIAM,

$$\triangle, M = \mathcal{T}(|F'_{t_1,i+1}(m,n) - F'_{t_2,i+1}(m,n)|_1), \tag{14}$$

$$(k,l) = (m,n) + (\triangle_x + \triangle_y), \tag{15}$$

$$\hat{F}_{t,i}(m,n) = BR(\sum w(k,l) M(k,l) F_{t,i}(k,l)) \tag{16}$$

where $\mathcal{T}$ is an implicit prediction function used to generate the offset values $\triangle \in \mathbb{R}$ and the corresponding modulation scalars $M \in \mathbb{R}$. w represents the projection weights. Compared with the local cross-attention operation in BIAM, deformable convolution is a more sparse operation to achieve a large OCS [52], thus saving computing resources. Besides, according to Eq. 14, the deformable-relevant parameters are learned from the neighbor-aligned difference features. This strategy helps aggregate the bitemporal features at key points that can mitigate misalignment noise. It is like the human mind focusing on some key regions to identify real changes. Notably, from Eqs. 4 and 5, DIAM is depended on the output of BIAM. The refined bitemporal features are calculated as,

$$F'_{t,i}(m,n) = F_{t,i}(m,n) + \hat{F}_{t,i}(m,n), \tag{17}$$

where $F_{t,i}(m,n)$ is obtained in Eq. 2.

IV. Experiments

In this section, the datasets employed and the implementation details of the proposed method are first given. Second, we conduct the ablation studies and discussion to demonstrate the effectiveness of our method. Then, the comparison analysis with advanced methods from three aspects: the quantitative results, the visual results, and the complexity is implemented to illustrate the superiority of the presented method.

A. Datasets and Implementation Details

Datasets: In this work, three well-recognized build change detection datasets are employed to discuss the effectiveness of the proposed method, i.e., LEVIR-CD+ [14], WHU-CD [53], and GZ-CD [54]. LEVIR-CD+ is an extended version of the dataset LEVIR-CD, where the size of the image is 1024× 1024 pixels. It includes some challenging examples with registration errors. There is an official setting about the training and the testing data, which is followed in our experiments. WHU-CD is an aerial image building change detection dataset. It consists of two large-scale images for

training and testing. Notably, some confusing annotations in the WHU-CD dataset are manually corrected in this work. GZ-CD is a large-scale very high-resolution remote sensing image building change detection dataset, where the paired bitemporal images are acquired during the periods between 2006 and 2019 and the image size ranges from 1006×1168 pixels to 4936×5224 pixels. For the sake of convenience in subsequent computing, the images in LEVIR-CD+ and WHU-CD datasets are cropped to 512×512 pixels for training and testing. Meanwhile, the images in GZ-CD are cropped to 256×256 pixels, thereby generating 2504 pairs of images for training and 626 pairs of images for testing.

Implementation Details: The batch size during training for the LEVIR-CD+, WHU-CD, and GZ-CD datasets is 4, 4, and 16 respectively. The total epochs are 200. The Adamw optimizer and the learning rate linear decay strategy are utilized in the proposed method. The initial learning rate is set to 0.0001. Meanwhile, the data augmentation techniques including random flip, rotate, gaussian blur, and color jitter are adopted during training. Besides, the OCS $\mathcal{R}$ in Eq. 12 is a 3×3 local grid centered as the observed pixel without special illustration, and the kernel size of the deformable convolution described from Eqs. 14 to 16 is also 3×3. The quantitative evaluation metrics employed to compare the performance of different methods are precision (P), recall (R), F1-score (F1), intersection of union (IoU), and overall accuracy (OA). All experiments are implemented on one NVIDIA 3090 GPU.

B. Ablation Studies and Discussion

TABLE I
THE ALATION STUDIES ON THE LEVIR-CD+ DATASET.
THE BEST VALUES ARE IN BOLD (%).

SAFM	BIAM	DIAM	P	R	F1	IoU	OA	IT(ms)
			85.25	82.12	83.65	71.90	98.69	36.13
✓			84.60	83.02	83.80	72.12	98.69	37.95
✓	✓		84.55	**85.36**	84.95	73.84	98.77	38.86
w/o DC	✓		85.05	84.31	84.68↓	73.43↓	98.76	39.85
	✓		86.58	82.71	84.60	73.31	98.78	34.83
	w/o g		84.29	83.41	83.85↓	72.19↓	98.69	35.26
	✓	✓	**87.09**	82.89	84.94	73.82	98.80	36.03
✓	✓	✓	86.03	84.30	**85.15**	**74.14**	**98.80**	39.38

TABLE II
THE ANALYSIS ON THE EFFECT OF THE BIAM ($\mathcal{R}/3$).
THE BEST VALUES ARE IN BOLD (%).

Baseline+BIAM				Evaluation Metrics				
S1	S2	S3	S4	P	R	F1	IoU	OA
✓				83.79	**83.92**	83.86	72.20	98.69
	✓			85.15	83.67	84.40	73.01	98.74
		✓		86.01	82.14	84.03	72.46	98.73
			✓	**86.58**	82.71	**84.60**	**73.31**	**98.78**

To verify the effectiveness of the CASP, we conduct extensive experiments to explore the effects of the designed modules and some parameter settings. Meanwhile, we discuss the impact of pretrained weights and backbone networks.

TABLE III
THE ANALYSIS ON THE EFFECT OF THE BIAM.
THE BEST VALUES ARE IN BOLD (%).

Baseline+BIAM			Evaluation Metrics					
S4	S3	S2	P	R	F1	IoU	OA	IT(ms)
$\mathcal{R}/3$			**86.58**	82.71	84.60	73.31	**98.78**	34.83
$\mathcal{R}/3$	$\mathcal{R}/5$		86.16	83.12	84.62	73.33	98.77	36.86
$\mathcal{R}/3$	$\mathcal{R}/5$	$\mathcal{R}/7$	84.41	**85.22**	**84.81**	**73.63**	98.76	42.01

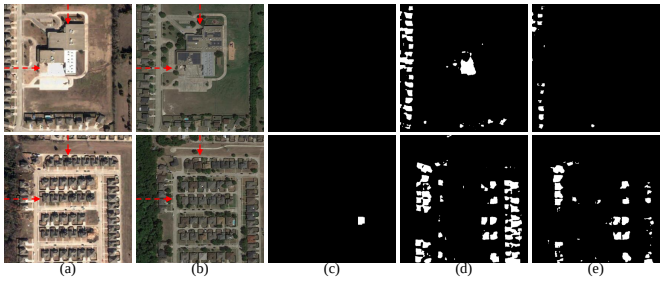

Fig. 6. The qualitative analysis of the BIAM. (a) I_{t1}. (b) I_{t2}. (c) Ground Truth. (d) Baseline. (c) Baseline+BIAM.

Ablation Studies: In this section, the backbone network used in the feature encoder is ResNet18, and the model variants for ablation analysis are trained from scratch. The experiments are performed on the LEVIR-CD+ dataset, where some bitemporal images are misaligned slightly. Therefore, its experimental results are more convincing than others in demonstrating the effects of the designed modules. Table. I reports the quantitative metrics.

Firstly, from Table. I, the F1/IoU can increase a bit after inserting the SAFM into the feature encoder. Meanwhile, it can be noted that the recall can be well improved after employing SAFM, which means more changed building regions are identified and indirectly indicates the enhancement of the representation ability of the building object. The negative effect is the decline of the precision, which can be attributed to some interference amplified, e.g., non-building objects of change. Besides, the importance of DC (Eq. 9) in SAFM is explored by replacing it with a vanilla convolution. In detail, from the 3-th and 4-th rows in Table. I, the F1/IoU reduce as expected without DC in SAFM. Overall, the SAFM is beneficial to the improvement of the model performance.

Secondly, the efficacy of the feature context alignment strategy including BIAM and DIAM is explored. From the 5-th and 7-th rows in Table. I, the injection of BIAM and DIAM improves the accuracy of the baseline model significantly. Meanwhile, the precision increases without the recall decline after introducing them. It can be inferred that they can effectively suppress the adverse impact of the slight misalignment error. In addition, the results in the 6-th row confirm the role of the guided term g (Eq. 12) in BIAM is important.

Furthermore, we validate the rationale for employing the BIAM exclusively in stage 4. First, it can inferred that the computational burden will increase when BIAM is employed in the other feature levels (i=1, 2, 3), owing to the increase in the feature size. Second, Table. II indicates that the model performs best when BIAM is applied to the deepest feature

layer, where the observed range corresponds to a larger range of the original image. Third, we explore the necessity of introducing the BIAM in multiple stages. As shown in Table. III, introducing the BIAM in stages 2 and 3 can slightly improve the F1/IoU. However, the computational cost is expensive and the OA is reduced. In contrast, the DIAM can help achieve better accuracy with fewer inference times (IT). Therefore, from the perspective of computational consumption and performance improvement, the BIAM is only employed in stage 4 in our model and the DIAM is integrated in stages 2 and 3. Finally, Fig. 6. gives the qualitative verification of BIAM, where the red arrow is used to indicate the spatial offset. It can be seen that BIAM effectively alleviates the false alarms caused by the content offset in bitemporal images.

TABLE IV
THE EFFECT OF THE PRETRAINED WEIGHTS (%).

Models	Pretrained	P	R	F1	IoU	OA
Baseline		85.25	82.12	83.65	71.90	98.69
	✓	85.14	85.11	85.13	74.10	98.79
CASP-R18		86.03	84.30	85.15	74.14	98.80
	✓	87.33	85.55	86.43	76.10	98.91

TABLE V
PERFORMANCE COMPARISON OF DIFFERENT BACKBONES. THE BEST VALUES ARE IN BOLD (%).

Backbones	LEVIR-CD+		WHU-CD		GZ-CD	
	F1	IoU	F1	IoU	F1	IoU
ResNet18	86.43	76.10	92.26	85.62	87.26	77.40
ResNet34	86.52	76.25	92.56	86.14	88.02	78.60
ResNet50	86.32	75.93	92.55	86.13	88.40	79.22
MobileNetv2	86.67	76.48	92.38	85.85	87.89	78.48
MiT-b0	86.96	76.93	93.32	87.47	88.19	78.87
MiT-b2	**87.32**	**77.49**	**93.95**	**88.59**	**89.18**	**80.47**

In summary, the SAFM can effectively reduce omissions, while the BIAM and DIAM can enhance precision. Consequently, the combined use of these modules significantly improves overall performance.

Discussion: Are the pretrained weights and the backbone networks key factors affecting the performance of the model? Firstly, as shown in Table. IV, the F1/IoU of the baseline model with the backbone network ResNet18 and the CASP-R18 are improved by a large margin after adopting the pretrained weights initialization for training, where the feature encoder (i.e., ResNet18) is pretrained on the ImageNet-1K. Secondly, Table. V reports the accuracy of the models with different feature encoders, i.e., backbone networks. The first four are CNN models, and the last two are Transformer models. Among them, the Transformer models are superior to the CNN models overall, and the gaps between the highest and lowest IoU are even as high as 1.56%, 2.97%, and 3.07% on the three datasets respectively. Therefore, there is no doubt that it is important to load pretrained weights during training and select an advanced backbone network for reaching competitive performance.

C. Comparison

To demonstrate the superiority of CASP, the eight latest representative models including DTCDSCN [55], SSCD [56], DGANet [15], DMINet [57], ICIF-Net [58], APD [33], ELGCNet [59], and BiFA [34] are compared with CASP. They are easily adapted to building change detection. Among them, DGANet is a metric-based model. APD and BiFA explore context alignment strategies to improve accuracy. For a fair comparison, we select our models using the ResNet18 and MiT-b0 as feature encoders compared to other methods.

TABLE VI
THE QUANTITATIVE COMPARISON ON THE LEVIR-CD+ DATASET. THE BEST VALUES ARE IN BOLD (%).

Methods	P	R	F1	IoU	OA
DTCDSCN	81.32	**86.62**	83.88	72.24	98.65
SSCD	84.47	83.54	84.00	72.42	98.71
DGANet	85.39	79.37	82.47	70.16	98.62
DMINET	85.47	81.76	83.58	71.79	98.69
ICIF-Net	85.26	81.33	83.25	71.30	98.67
APD	85.71	82.76	84.21	72.72	98.74
ELGCNet	86.50	82.82	84.62	73.34	98.78
BiFA	82.70	83.47	83.08	71.06	98.62
CASP-R18	87.33	85.55	86.43	76.10	98.91
CASP-Mb0	**87.34**	86.59	**86.96**	**76.93**	**98.94**

Results on LEVIR-CD+: This dataset contains many samples with slight geometric distortion. Consequently, this comparison can effectively demonstrate the efficacy and superiority of our solution. Table. VI reports the quantitative results. The qualitative comparison is shown in Fig. 7(a). From them, the performance of CASP-R18 and CASP-Mb0 is far superior to the compared methods. Although the recall of DTCDSCN is the highest, its precision is the lowest. It can be found that there are many false positives predicted by it from the second sample presented in Fig.7(a), which is consistent with the quantitative metrics. APD and BiFA introduce the bitemporal alignment mechanisms. However, their accuracy is inferior to ours. Specifically, the visual results indicate that APD is ineffective in suppressing misalignment noise, resulting in numerous false detections. Besides, from the given samples, it can be noted that DGANet alleviates the adverse effect of registration errors well. However, the number of its omissions is much higher than other methods. It can be attributed to its high threshold for change identification, which is beneficial to achieve high precision but sacrifices recall. The suitable threshold for DGANet is hard to decide. Without extra parameter settings during testing, CASP-R18 and CASP-Mb0 achieve high precision with high recall, and the interference of geometric dislocation is well suppressed.

Results on WHU-CD: WHU-CD is an aerial image building change detection dataset. The bitemporal images in it are well aligned, but the high-resolution images bring the challenge of significant intra-class diversity and inter-class similarity, such as the building roof and the road surface. Therefore, it is important to extract discriminative building representation for accurate change identification. Intuitively, the structure perception mechanism attentively introduced in our method is helpful to improve accuracy. The quantitative

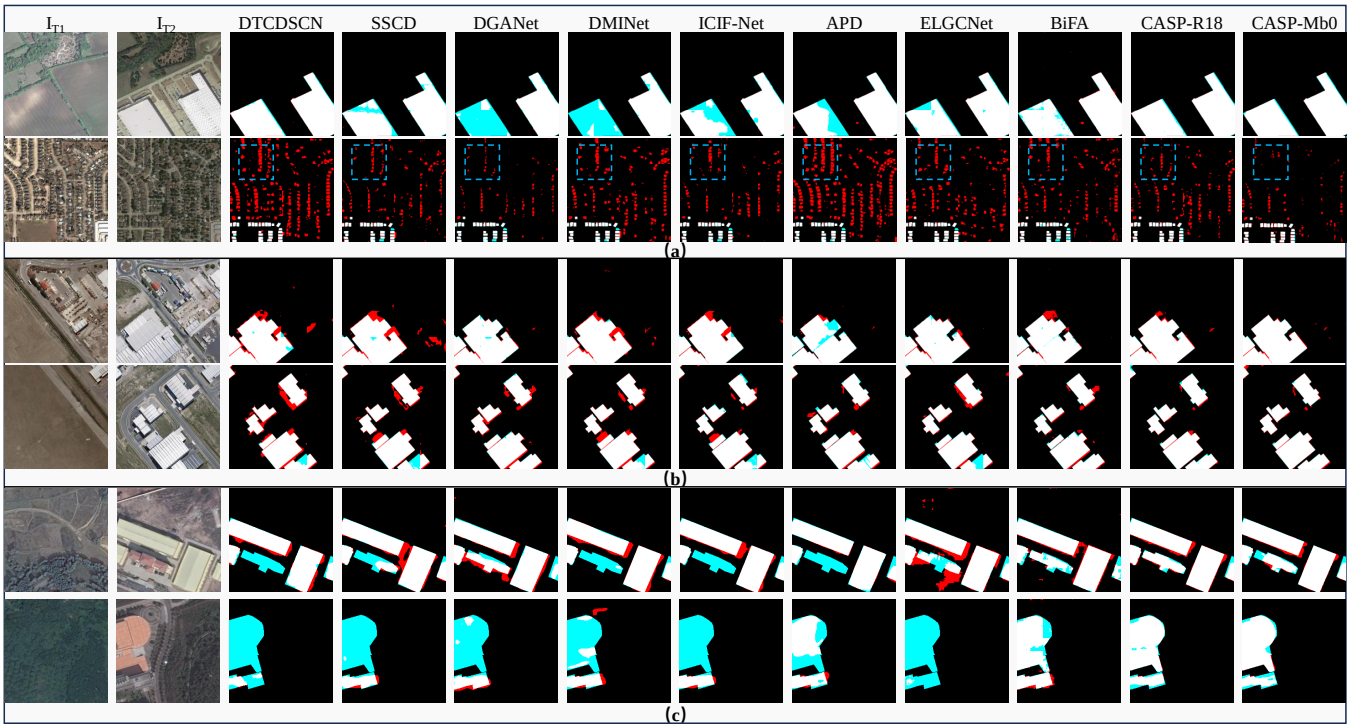

Fig. 7. The qualitative visual comparison on three datasets. (a) LEVIR-CD+. (b) WHU-CD. (c) GZ-CD. The predicted true positives, false positives, false negatives, and true negatives are presented in white, red, blue, and black respectively.

TABLE VII
THE QUANTITATIVE COMPARISON ON THE WHU-CD DATASET. THE BEST VALUES ARE IN BOLD (%).

Methods	P	R	F1	IoU	OA
DTCDSCN	71.15	89.39	79.24	65.61	98.29
SSCD	82.32	87.26	84.72	73.49	98.85
DGANet	94.58	86.72	90.48	82.61	99.34
DMINET	80.31	85.04	82.61	70.37	98.70
ICIF-Net	88.43	80.89	84.49	73.14	98.92
APD	94.25	85.16	89.47	80.95	99.27
ELGCNet	91.48	85.00	88.12	78.77	99.17
BiFA	94.70	90.19	92.39	85.86	99.46
CASP-R18	**96.26**	88.57	92.26	85.62	99.46
CASP-Mb0	95.80	**90.96**	**93.32**	**87.47**	**99.53**

TABLE VIII
THE QUANTITATIVE COMPARISON ON THE GZ-CD DATASET. THE BEST VALUES ARE IN BOLD (%).

Methods	P	R	F1	IoU	OA
DTCDSCN	83.82	83.69	83.75	72.05	97.27
SSCD	87.05	80.53	83.67	71.92	97.35
DGANet	88.95	82.57	85.64	74.88	97.67
DMINET	86.55	83.11	84.80	73.61	97.49
ICIF-Net	85.26	84.11	84.68	73.43	97.44
APD	89.61	82.42	85.87	75.32	97.71
ELGCNet	84.65	80.95	82.76	70.59	97.16
BiFA	**90.86**	**86.13**	**88.43**	**79.26**	**98.10**
CASP-R18	89.04	85.55	87.26	77.40	97.90
CASP-Mb0	90.60	85.90	88.19	78.87	98.06

results reported in Table. VII indirectly indicate this point. The performance of CASP-Mb0 is the best. The next best is BiFA. Notably, its backbone is MiT-b0, being the same as CASP-Mb0. To be fair, the F1 of CASP is 0.93% higher than its. Meanwhile, the F1 of our CASP-R18 are almost consistent with that of BiFA. Compared to other methods excluding BiFA, the superiority of CASP is significant. The Fig. 7(b) shows the visual results. From it, the edge details of CASP are more refined than others. Besides, the integrity of detected results is higher, and the errors are less.

Results on GZ-CD: GZ-CD is a small-scale satellite image building change detection dataset. Similar to the WHU-CD, the bitemporal images in it are well co-registered. The main challenge is the limited training samples and the intra-class variation between instances. Table. VIII reports the quantitative metrics. BiFA achieves the SoTA performance. CASP-Mb0 is the suboptimal. Their performance is close, where the F1 of CASP-Mb0 is only 0.24% less than that of BiFA. It can be attributed to the small size of the GZ-CD dataset. The BiFA possesses higher parameters and computational complexity, thus it achieves better performance than our method within acceptable ranges. However, its generalization ability is far inferior to that of CASP, which is reported in Table. IX. Besides, compared to the methods using CNNs as the feature encoder, such as DGANet, SSCD, and APD, CASP-R18 is superior to them. According to the five metrics reported in Table. VIII, it can be concluded that the performance of our method is comparable. The qualitative comparison is given in Fig. 7(c). It can be noted that an instance in the second example is missed by most methods, which can be attributed to its special color and shape. Meanwhile, the visual results show the false positives of the proposed method are less

at the boundary regions, and the predicted integrity is also competitive.

To sum up, we compare the performance of CASP with multiple excellent and latest methods on three datasets, from aerial images to satellite images, from the misaligned images to the well co-registered images. Overall, CASP is competitive and further improves the accuracy of building change detection.

TABLE IX
THE CROSS-DOMAIN TESTING ACROSS THE GZ-CD DATASET AND THE WHU-CD DATASET. (GZ-CD → WHU-CD).
THE BEST VALUES ARE IN BOLD (%).

Methods	P	R	F1	IoU	OA
BiFA	37.44	66.15	47.82	31.44	94.74
CASP-Mb0	**64.27**	**73.25**	**68.46**	**52.05**	**97.54**

TABLE X
THE COMPLEXITY COMPARISON OF DIFFERENT METHODS.

Method	Params(M)	FLOPs(G)	F1(%)	IoU(%)
DTCDSCN	41.07	13.21	83.88	72.24
SSCD	12.17	9.55	84.00	72.42
DGANet	12.28	12.56	82.47	70.16
DMINet	6.76	14.55	83.58	71.79
ICIF-Net	25.83	25.41	83.25	71.30
APD	110.90	23.37	84.21	72.72
ELGCNet	10.57	187.98	84.62	73.34
BiFA	9.87	53.00	83.08	71.06
CASP-R18	14.55	9.19	86.43	76.10
CASP-Mb0	4.57	2.63	86.96	76.93

Complexity: In this section, we discuss the complexity of the proposed method and further illustrate its superiority. Table. X reports the complexity metrics Params/FLOPs, and the performance metrics F1/IoU of compared methods on the LEVIR-CD+ dataset. CASP-Mb0 achieves the best performance with the lowest FLOPs and the fewest Params. Moreover, the FLOPs of the CASP-R18 are relatively competitive. Although the Params of the DMINet are few, its performance reported on the three datasets is far inferior to the CASP. The results show the accuracy of the BiFA on the WHU-CD and the GZ-CD datasets is competitive, but its FLOPs are the second-highest. Overall, compared to the eight methods, CASP is efficient and competitive.

V. CONCLUSION

In this paper, we explore the importance of context alignment and structure perception for accurately detecting building changes in the presence of slight misregistration in bitemporal images. First, a neighbor feature fusion module integrating the differential convolution is designed to generate edge-structure-enhanced features, which help enhance the discrimination of building representation and refine the change maps. Second, a bitemporal interactive alignment module is developed, which investigates a bidirectional reference-guided aggregation strategy to imitate humans to identify changes, aiming at modulating misaligned context information. It effectively improves the performance of the baseline model. Meanwhile, relying on its output, a difference-induced alignment module with deformable convolution further helps to boost the immunity to spatial offsets. Compared to some advanced change detection models, the accuracy of the proposed method is far superior to theirs in the LEVIR-CD+ dataset, a typical dataset with misaligned bitemporal images. Meanwhile, it is competitive on the other two datasets. In the future, alignment strategies for handling complex off-nadir misalignment caused by view differences, not slight offsets will be explored, advancing change detection models towards practical applications.

REFERENCES

[1] N. Quarmby and J. Cushnie, "Monitoring urban land cover changes at the urban fringe from spot hrv imagery in south-east england," *International Journal of Remote Sensing*, vol. 10, no. 6, pp. 953–963, 1989.

[2] 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 Sensing of Environment*, vol. 265, p. 112636, 2021.

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

[4] C. Wu, L. Zhang, B. Du, H. Chen, J. Wang, and H. Zhong, "Unet-like remote sensing change detection: A review of current models and research directions," *IEEE Geoscience and Remote Sensing Magazine*, pp. 2–31, 2024.

[5] R. Johnson and E. Kasischke, "Change Vector Analysis: A technique for the multispectral monitoring of land cover and condition," *International Journal of Remote Sensing*, vol. 19, no. 3, pp. 411–426, 1998.

[6] T. Habib, J. Inglada, G. Mercier, and J. Chanussot, "Support vector reduction in svm algorithm for abrupt change detection in remote sensing," *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 3, pp. 606–610, 2009.

[7] J. Deng, K. Wang, Y. Deng, and G. Qi, "PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data," *International Journal of Remote Sensing*, vol. 29, no. 16, pp. 4823–4838, 2008.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[9] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[10] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.

[11] W. Bandara and V. Patel, "A transformer-based siamese network for change detection," in *Proc. IEEE International Geoscience and Remote Sensing Symposium*, 2022, pp. 207–210.

[12] C. Zhang, L. Wang, S. Cheng, and Y. Li, "SwinSUNet: Pure transformer network for remote sensing image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[13] Q. Li, R. Zhong, X. Du, and Y. Du, "TransUNetCD: A hybrid transformer network for change detection in optical remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–19, 2022.

[14] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sensing*, vol. 12, no. 10, p. 1662, 2020.

[15] M. Zhang, Q. Li, Y. Miao, Y. Yuan, and Q. Wang, "Difference-guided aggregation network with multiimage pixel contrast for change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[16] S. Fang, K. Li, and Z. Li, "Changer: Feature interaction is what you need for change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–11, 2023.

[17] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[18] S. Tian, X. Tan, A. Ma, Z. Zheng, L. Zhang, and Y. Zhong, "Temporal-agnostic change region proposal for semantic change detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 204, pp. 306–320, 2023.

[19] K. Zhang, X. Zhao, F. Zhang, L. Ding, J. Sun, and L. Bruzzone, "Relation changes matter: Cross-temporal difference transformer for change detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[20] Y. Liu, F. Zhang, S. Zhang, K. Zhang, J. Sun, and L. Bruzzone, "Content-guided spatial–spectral integration network for change detection in hr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024.

[21] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected siamese network for change detection of vhr images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[22] Z. Li, C. Yan, Y. Sun, and Q. Xin, "A densely attentive refinement network for change detection based on very-high-resolution bitemporal remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2022.

[23] G. Wang, G. Cheng, P. Zhou, and J. Han, "Cross-level attentive feature aggregation for change detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[24] L. Wang, J. Zhang, and L. Bruzzone, "Mixcdnet: A lightweight change detection network mixing features across cnn and transformer," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[26] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection." in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 936–944.

[27] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[28] M. Zhang, Q. Li, Y. Yuan, and Q. Wang, "Edge neighborhood contrastive learning for building change detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[29] A. Eftekhari, F. Samadzadegan, and F. D. Javan, "Building change detection using the parallel spatial-channel attention block and edge-guided deep network," *International Journal of Applied Earth Observation and Geoinformation*, vol. 117, p. 103180, 2023.

[30] H. Du, Z. Huang, and Y. Zhang, "An optimized edge-focused siamese network for monitoring new illegal buildings using satellite images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[31] I.-H. Lee and T.-S. Choi, "Accurate registration using adaptive block processing for multispectral images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 9, pp. 1491–1501, 2013.

[32] J.-M. Park, U.-H. Kim, S.-H. Lee, and J.-H. Kim, "Dual task learning by leveraging both dense correspondence and mis-correspondence for robust change detection with imperfect matches," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 749–13 759.

[33] S. Wang, Y. Li, M. Xie, M. Chi, Y. Wang, C. Wang, and W. Zhu, "Align, perturb and decouple: toward better leverage of difference information for rsi change detection," in *Proc. International Joint Conference on Artificial Intelligence*, 2023.

[34] H. Zhang, H. Chen, C. Zhou, K. Chen, C. Liu, Z. Zou, and Z. Shi, "BiFA: Remote sensing image change detection with bitemporal feature alignment," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[35] Y. Zhao, H.-C. Li, N. Liu, and R. Wang, "Toward distortion-aware change detection in realistic scenarios," *arXiv preprint arXiv:2401.05157*, 2024.

[36] J. Tian, D. Peng, H. Guan, and H. Ding, "RACDNet: Resolution-and alignment-aware change detection network for optical remote sensing imagery," *Remote Sensing*, vol. 14, no. 18, p. 4527, 2022.

[37] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE International Conference on Computer Vision*, 2017, pp. 764–773.

[38] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9308–9316.

[39] Y. Liu, K. Zhang, C. Guan, S. Zhang, H. Li, W. Wan, and J. Sun, "Building change detection in earthquake: A multi-scale interaction network with offset calibration and a dataset," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[40] Y. Liang, X. Xu, C. Zhang, J. Liu, D. Wang, and M. Han, "Progressive difference amplification network with edge sensitivity for remote sensing image change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[41] M. Yang, Y. Zhou, Y. Feng, and S. Huo, "Edge-guided hierarchical network for building change detection in remote sensing images," *Applied Sciences*, vol. 14, no. 13, p. 5415, 2024.

[42] C. Yang, M. Chen, Z. Xiong, Y. Yuan, and Q. Wang, "CM-Net: Concentric mask based arbitrary-shaped text detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 2864–2877, 2022.

[43] Y. Zhu, K. Lv, Y. Yu, and W. Xu, "Edge-guided parallel network for vhr remote sensing image change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 7791–7803, 2023.

[44] J. Ma, J. Duan, X. Tang, X. Zhang, and L. Jiao, "EATDer: Edge-assisted adaptive transformer detector for remote sensing change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[45] B. Bai, W. Fu, T. Lu, and S. Li, "Edge-guided recurrent convolutional neural network for multitemporal remote sensing image building change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[46] Z. Chen, Y. Zhou, B. Wang, X. Xu, N. He, S. Jin, and S. Jin, "EGDE-Net: A building change detection method for high-resolution remote sensing imagery based on edge guidance and differential enhancement," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 191, pp. 203–222, 2022.

[47] H. Du, Z. Huang, and Y. Zhang, "An optimized edge-focused siamese network for monitoring new illegal buildings using satellite images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[48] A. Eftekhari, F. Samadzadegan, and F. Dadrass Javan, "Building change detection using the parallel spatial-channel attention block and edge-guided deep network," *International Journal of Applied Earth Observation and Geoinformation*, vol. 117, p. 103180, 2023.

[49] Q. Chen, Q. Wu, J. Wang, Q. Hu, T. Hu, E. Ding, J. Cheng, and J. Wang, "Mixformer: Mixing features across windows and dimensions," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5249–5259.

[50] M. Zhang, Q. Li, Y. Yuan, and Q. Wang, "Boosting binary object change detection via unpaired image prototypes contrast," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–9, 2024.

[51] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao, "Searching central difference convolutional networks for face anti-spoofing," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5295–5305.

[52] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li *et al.*, "Internimage: Exploring large-scale vision foundation models with deformable convolutions," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 408–14 419.

[53] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 574–586, 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 Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 5891–5906, 2021.

[55] 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 Geoscience and Remote Sensing Letters*, vol. 18, no. 5, pp. 811–815, 2021.

[56] L. Wang, Y. Fang, Z. Li, C. Wu, M. Xu, and M. Shao, "Summator–Subtractor Network: Modeling spatial and channel differences for change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–12, 2024.

[57] Y. Feng, J. Jiang, H. Xu, and J. Zheng, "Change detection on remote sensing images using dual-branch multilevel intertemporal network,"

IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–15, 2023.

[58] Y. Feng, H. Xu, J. Jiang, H. Liu, and J. Zheng, "ICIF-Net: Intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[59] M. Noman, M. Fiaz, H. Cholakkal, S. Khan, and F. S. Khan, "ELGC-Net: Efficient local–global context aggregation for remote sensing change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–11, 2024.

Qi Wang (Senior Member, IEEE) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN). Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, pattern recognition and remote sensing. For more information, visit the link (https://crabwq.github.io/).

Mingwei Zhang received the B.E. degree in automation from Zhengzhou University, Zhengzhou, China, in 2021, and the M.S. degree from the Unmanned System Research Institute, Northwestern Polytechnical University, Xi'an, China, in 2024. He is currently pursuing the Ph.D. degree with the School of Computer Science, Northwestern Polytechnical University, Xi'an, China. His research interests include remote sensing image acquisition and processing.

Jiawei Ren received his B.E. degree in software engineering from the North University of China, Shanxi Province, China in 2023. He is currently pursuing a M.S. degree at the School of Software, Northwestern Polytechnical University. His research interests include image processing, deep learning and computer vision.

Qiang Li received the Ph.D. degree in computer science and technology from Northwestern Polytechnical University, Xi'an, China in 2022. He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include remote sensing image processing and computer vision.