# Difference-Guided Aggregation Network with Multi-Image Pixel Contrast for Change Detection

Mingwei Zhang, Qiang Li, *Member, IEEE*, Yanling Miao, Yuan Yuan, *Senior Member, IEEE*, and Qi Wang, *Senior Member, IEEE*

***Abstract*—Change detection is a critical task in remote sensing to monitor the state of the surface on Earth. This field has been dominated by deep learning-based methods recently. Many models that model the temporal-spatial correlation in bitemporal images through the non-local interaction between bitemporal features achieve impressive performance. However, under complex scenes including multiple change types or weakly discriminate objects, they suffer from achieving discriminative fusion of information due to the weak semantic discrimination of the bitemporal representations. Aiming at this problem, a difference-guided aggregation network (DGANet) is proposed, where two key modules are injected, i.e., a difference-guided aggregation module (DGAM) and a weighted metric module (WMM). The bitemporal features in DGAM are aggregated with the guidance of their differences, which focuses on their change relevance and relaxes their semantic distinction. Therefore, the fused features are change-relevant and discriminative. WMM aims to achieve adaptive distance computation between the bitemporal features by dynamic feature attention in different dimensions. It is helpful to suppress the pseudo-changes. Besides, a change magnitude contrastive loss (CMCL) is introduced to employ the dependency of bitemporal pixels in different bitemporal images, which further enhances the representation quality of the model. Meanwhile, it is further extended in this work. The effectiveness of the three improvements is demonstrated by extensive ablation studies. The results on three datasets widely used illustrate that our method achieves satisfactory performance.**

***Index Terms*—Change detection, difference-guided aggregation, adaptive metric, contrastive loss.**

## I. Introduction

CHANGE detection is a hot topic in the field of remote sensing. It aims to identify the differences between two co-registered remote sensing images obtained in an identical region. It has been applied in many fields including urban planning [1], disaster assessment [2], etc.

In the early days, an independent pixel is regarded as the basic processing unit, i.e., pixel-based change detection (PBCD). PBCD algorithms can be divided into three types simply including algebra-, transformation-, and classification-based methods. Algebra-based methods usually obtain a difference image by mathematical algebra operations between the bitemporal images to locate the changed regions. The algebra operations consist of image ratio [3], image difference [4], [5], change vector analysis (CVA) [6], etc. Transform-based methods implement the change analysis on an effective feature space, which is translated from original images by some pattern analysis approaches, such as principal component analysis [7] and random forest regression [8]. Classification-based methods decide the state of the pixels by comparing the pixel-wise classification results of the bitemporal images, whose accuracy highly depends on the performance of the utilized classifier [9]. PBCD is sensitive to noise and not suitable for high-resolution remote sensing images with rich details. Besides, PBCD ignores usable contextual information around the individual pixel. Later, the pixels are not individually processed but combined, and the low-level image features containing spectral, texture, and shape are employed [10], [11]. Nevertheless, interfered easily by some adverse factors, such as illumination differences, season variations, and registration errors, handcrafted features are prone to pseudo-change detection results.

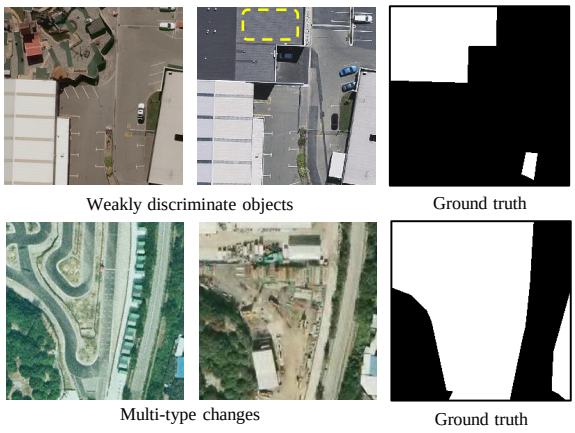



Fig. 1. The illustration of challenging scenarios including multi-type changes and weakly discriminate objects. The characteristic of the labeled building presented in the first example is similar to that of the ground. The second example shows cluttered multi-type objects in changed regions.

Recently, there are many models based on deep learning proposed for change detection. Supervised change detection models are discussed here. They are summarized into two types: classification- and metric-based approaches. Classification-based models [12]–[15] aim to predict the change confidence of every pixel. Daudt *et al.* [16] explore three strategies including image concatenation, feature differ-

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

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

Qiang Li, Yuan Yuan, and Qi Wang are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P.R. China (e-mail: liqmges@gmail.com, y.yuan1.ieee@gmail.com, crabwq@gmail.com) (*Corresponding author: Qi Wang, Qiang Li).*

ence, and feature concatenation based on the full convolutional network (FCN) for change information extraction. It ignores the temporal relation in bitemporal images. To solve this problem, Papadomanolaki *et al.* [17] integrate the long short-term memory networks (LSTMs) into a U-Net architecture to extract discriminative features and model the temporal dependency jointly. Besides, some methods introduce extra information to assist change detection. Chen *et al.* [18] employ the edge of changed regions to guide the network preserving the structures of changed objects, which alleviates the loss of boundary information effectively. Metric-based models identify the categories of the pixels by measuring the distance between the features extracted from the bitemporal images [19]–[24]. Chen *et al.* [25] propose a spatial-temporal attention network to fuse multi-scale features, which effectively model the spatial interdependence in bitemporal images. To alleviate the effect of pseudo-changes, Shi *et al.* [26] introduce a deeply supervised strategy and an attention mechanism in their model. Notably, the metric- and classification-based models discussed below are all supervised and defined as introduced above.

From recent advances, the main obstacles in change detection are well mitigated, e.g., the acquisition of representative features and the suppression of pseudo-changes. In particular, many methods employ a spatial-temporal attention mechanism (STAM) to achieve impressive performance [27]–[29], where STAM effectively models the temporal-spatial correlation in bitemporal images by the non-local interaction between the bitemporal features. However, STAM requires bitemporal representations with well-semantic discrimination. Otherwise, as shown in Fig. 1, when the semantic distinction among representations in interested changed regions is weak, land covers are easily confused in representations generated by it. Accurate change detection is hindered instead. Motivated by this challenging issue, a difference-guided aggregation network (DGANet) is proposed. It is a metric-based model comprising two core modules: a difference-guided aggregation module (DGAM) and a weighted metric module (WMM). Different from most methods, DGANet explores the interaction between the difference features and bitemporal features. It can acquire representative features even if the semantic distinction of the bitemporal embedding is weak. Besides, to further enhance the representation quality of the model, a change magnitude contrastive loss (CMCL) is introduced to investigate the correlation between bitemporal pixels in multiple bitemporal images. Notably, it is further extended in this work. It aims to make the representations extracted from bitemporal pixels with the same class (i.e., either change or non-change) more compact and the representations of those with different classes more discriminative. The main contributions are given in detail as follows:

- DGAM−It is an innovative aggregation approach for bitemporal features. In detail, feature differences are used to guide the global fusion of the bitemporal representations, which relaxes semantic distinction in them and enhances the change-relevance of fused features. Besides, a modulation mechanism in it is designed to make fused representations well adapted to metric-based change detection.
- WMM−Focusing on the state characteristic of the bitemporal pixels, it learns multi-dimension weights to dynamically maintain or reduce the errors of the bitemporal representations in different channels and spatial positions. Thus, it can achieve adaptive distance computation between bitemporal features and is helpful to suppress pseudo-changes.
- CMCL−It is a supervised contrastive loss introduced to explore the interdependence of bitemporal pixels in different geographical areas. We extend it so that it can be equipped with different types of supervised change detection models including metric- and classification-based models. Besides, for metric-based models, an alternative to CMCL is given.

The remainder of this article is organized as follows. Section II briefly discusses the related works. The proposed method is described in Section III. The datasets and experimental results are given in Section IV, where the effectiveness of the proposed method is demonstrated and discussed. Finally, this article is concluded in Section V.

## II. Related Work

In this section, the related works about the attention mechanism and contrastive learning are introduced briefly.

### A. Attention Mechanism

The attention mechanism is important in visual information acquisition for humans. Depending on it, humans can focus on where is informative within the holistic view rapidly. Motivated by this meaningful finding, the attention mechanism has been investigated and utilized exhaustively over the past decades [30]–[33].

The common attention mechanisms in modern convolutional neural networks can be summarized into three types roughly: self-attention, spatial attention, and channel attention [34]–[37]. Many remote sensing change detection methods introducing them demonstrates their importance for improving the accuracy of change identification [38], [39]. For example, IFN [40] employs the convolutional block attention module (CBAM) [35] to modulate the bitemporal features and the difference features, which reduces the uneven prediction of boundary structure and enhance the completeness of predicted changed objects. CBAM is also integrated into the DSAM-Net [26] for providing highly discriminative features, which effectively suppresses pseudo-changes. In particular, CBAM includes two sub-modules, a spatial attention module (SAM) and a channel attention module (CAM). Inspired by these advances, we design a module that can employ learnable multi-dimension attention weights to discriminably measure the differences between bitemporal features.

The self-attention mechanism usually is used to construct the STAM. STAM models the temporal-spatial dependency in bitemporal images, thus generating representative representations. In detail, a pyramid STAM is proposed in STANet [25] to refine the features of objects and alleviate the adverse influence of the registration error. DARNet [28] integrates a hybrid attention module that combines a STAM with a channel attention mechanism to attain discriminative features. Besides, Zhou *et al.* [27] employ a multi-head STAM to address the interference caused by different imaging conditions

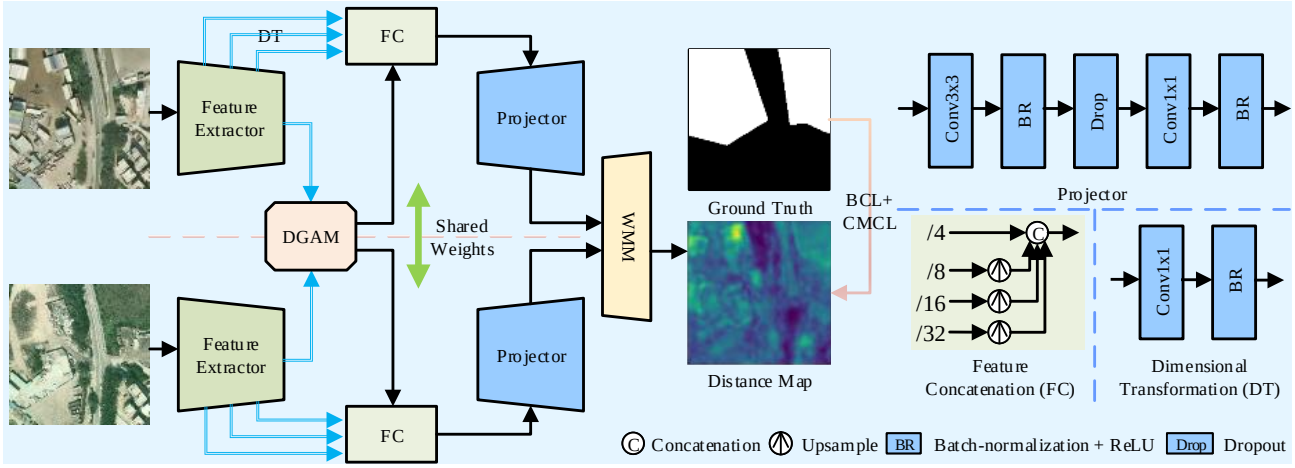



Fig. 2. The overall pipeline of the proposed algorithm for change detection.

in bitemporal images, where the inter-temporal information is used effectively. Encouraged by them, a difference-guided aggregation approach of bitemporal features is designed.

### B. Contrastive Learning

Contrastive learning has become a powerful tool for self-supervised representation learning (e.g. InstDisc [41], Moco [42], SimCLR [43], SwAV [44]). These methods use image-level instance discrimination as the pretext task, and then train networks to learn effective embedding without giving any labels by attracting the positive samples together and pushing the negative samples apart. The obtained visual representations show excellent performance on multiple vision tasks. In addition to the instance-level discrimination, some efforts employ pixel-level or patch-level comparison to learn the embedding that can be transferred to the downstream dense-prediction tasks well [45]. Chaitanya *et al.* [46] present a local contrastive loss, which aims to enhance the similarity of local regions with different views and repulsing other regions. Similarly, the invariance of local representations obtained through different transformations is utilized in DenseCL [47]. Xie *et al.* [48] take into account the pixel-to-propagation consistency. These methods show better results on semantic segmentation and object detection tasks than instance-level pretext tasks.

Some studies utilize known annotations to perform contrastive learning, i.e. supervised contrastive learning (SCL) [49]. Naturally, given a sample, SCL regards the samples belonging to the identical class as its positives and that from other categories as its negatives. Especially, when SCL is used for semantic segmentation, the samples attributed to the same class may be from multiple images. Therefore, the cross-image pixel relation can be explored and mined by SCL easily. A fully supervised pixel-level contrastive paradigm according to labeled data is proposed in [50]. Zhou *et al.* [51] extend it to weakly supervised segmentation. Inspired by these advances, SCL is adopted to investigate the association of bitemporal pixels in different bitemporal images. In this work, we further explore it by introducing and extending the contrastive loss in [52], where the association has been simply utilized for the edge neighborhood contrast.

## III. Methodology

In this section, the overall framework of the proposed model is first summarized. Then, we introduce and extend the CMCL. Finally, the complete structure of the DGAM and that of the WMM are described.

### A. Framework

The pipeline of our method is presented in Fig. 2, where the structure of the DGANet and the adopted loss functions are shown in detail. In addition, we give the inference process of the DGANet in Algorithm. 1. Firstly, multi-level embedding ($F_i^j, i = \{1, 2\}, j = \{1, 2, 3, 4\}$) is extracted by the feature extractor (FE) from given bitemporal images ($I_{T1}, I_{T2}$). To reduce the computational complexity of subsequent processes, the multi-level features are transformed into the same dimension by the operation of dimensional transformation (DT). It contains one $1\times1$ convolutional operation, one batch-normalization layer, and one activation function (i.e. $Conv$1x1+$BR$). Thereafter, the highest-level features with the lowest size are processed by the DGAM, which avoids bringing infeasible computational consumption under limited memory resources. The above operations aim to produce discriminative multi-level features. Among them, the low-level features own a high resolution and abundant details. The correlation of the bitemporal images is effectively modeled in the deepest-level features. Next, they are upsampled to the same size and concatenated for projection. The elements of the projector are shown in Fig. 2, where $Conv$3x3 represents the $3\times3$ convolutional layer and $Drop$ indicates the dropout layer. At last, the bitemporal outputs of the projector are fed into the WMM to obtain a distance map.

The model is updated by optimizing the batch-balanced contrastive loss (BCL) [25] and the CMCL. The BCL supervises the prediction of each pair of bitemporal pixels independently. Meanwhile, the CMCL models the dependency among them. Let $D \in \mathbb{R}^{H\times W}$ denote the distance map, where the value $d_{i,j}$ of every pixel is the distance between the bitemporal features calculated in the WMM. $i$ and $j$ represents the $i$-th row and $j$-th column. Let $Y \in \mathbb{R}^{H\times W}$ indicate the ground truth, where

**Algorithm 1:** Inference Process of DGANet

**Input:** A pair of geo-registered bitemporal images, $(I_{T1}, I_{T2})$
**Output:** A distance map, $D$

**for** $i \in \{1, 2\}$ **do**
  $F_i^1, F_i^2, F_i^3, F_i^4 \leftarrow \mathrm{FE}(I_{Ti})$;
**end**
**for** $i \in \{1, 2\}$ **do**
  **for** $j \in \{1, 2, 3, 4\}$ **do**
    $F_i^j \leftarrow \mathrm{DT}_j(F_i^j)$;
  **end**
**end**
$F_1^4, F_2^4 \leftarrow \mathrm{DGAM}(F_1^4, F_2^4)$;
**for** $i \in \{1, 2\}$ **do**
  $F_{Ti} \leftarrow \mathrm{FC}(F_i^1, F_i^2, F_i^3, F_i^4)$;
  $F_{Ti} \leftarrow \mathrm{Projector}(F_{Ti})$;
**end**
$D \leftarrow \mathrm{WMM}(F_{T1}, F_{T2})$;

$y_{i,j}$ is equal to 0 or 1. 0 and 1 refer to the non-change class and the change class. Given the distance maps and the ground truth within a mini-batch, the BCL can be computed as

$$\mathcal{L}^{BCL} = \frac{1}{2}\Big(\frac{1}{|\mathcal{B}^c|}\sum y_{i,j}\max(\Gamma - d_{i,j}, 0)^2 + \frac{1}{|\mathcal{B}^{\bar{c}}|}\sum(1 - y_{i,j})d_{i,j}^2\Big), \tag{1}$$

where $\mathcal{B}^c$ denotes the set of changed bitemporal pixels and $\mathcal{B}^{\bar{c}}$ represents the set of unchanged bitemporal pixels. The whole loss function is

$$\mathcal{L}^{CD} = \mathcal{L}^{BCL} + \gamma\mathcal{L}^{CMCL}, \tag{2}$$

where $\gamma$ is the tuned coefficient.

### B. Change Magnitude Contrastive Loss

Recent studies show that exploring the cross-image pixel relation benefits the semantic segmentation of natural images by supervised contrastive learning, where the semantic label of every pixel is given [30], [51]. A similar way can be used for change detection. In detail, according to the given state label of the bitemporal pixels, the relationship among bitemporal pixels in multiple bitemporal images is explored. We introduce a contrastive loss presented in [52], which is named the change magnitude contrastive loss (CMCL) here. It is formulated as

$$\mathcal{L}^{CMCL} = \frac{1}{|\mathcal{M}|}\sum_{k\in\mathcal{M}}\max\Big(\frac{1}{|\mathcal{P}^k|}\sum_{k^+\in\mathcal{P}^k} e_{k,k^+} - \frac{1}{|\mathcal{N}^k|}\sum_{k^-\in\mathcal{N}^k} e_{k,k^-} + \tau, 0\Big), \tag{3}$$

$$e_{k,k^+} = |m_k - m_{k^+}|, e_{k,k^-} = |m_k - m_{k^-}|, \tag{4}$$

where $\mathcal{P}^k$ and $\mathcal{N}^k$ represent the collections of positive samples and negative samples of the sample $k$ respectively. In particular, $\mathcal{M}$ indicates the collection of samples selected from multiple bitemporal images, where the samples of every class are composed of hard samples and easy samples. An incorrectly predicted sample is considered a hard sample. $m_k$ denotes the change magnitude of the sample $k$. The CMCL is initially tailored for metric-based models, where the distance between the bitemporal features is regarded as the change magnitude of the corresponding sample, i.e., $m_k = d_k$. Since the norm $h_k$ of the change representations in metric-based models is equal to the distance between the bitemporal features $d_k$, it can be got that $m_k = h_k$.

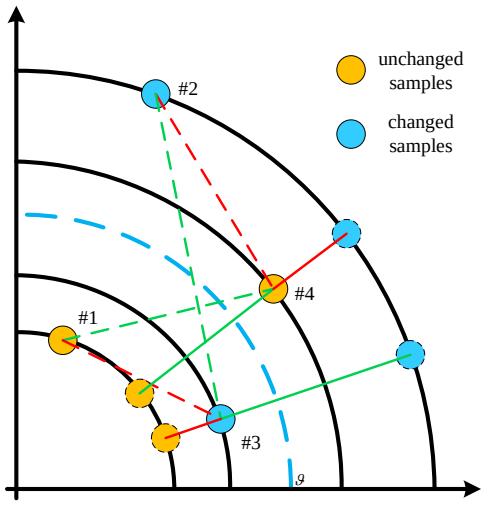



Fig. 3. The interpretable schematic of the change magnitude contrastive loss in the 2-D Euclidean space. The dashed line indicates the Euclidean distance between samples, where the green denotes that between samples of the same category and the red indicates that between samples with different classes. Correspondingly, the green and red solid lines refer to the absolute error of the change magnitude between different samples. Notably, sample #3 and sample #4 are wrongly predicted.

For metric-based models, we explore the basic mechanism of the CMCL in the 2-D Euclidean space (Fig. 3), where a 2-D vector is used to indicate the change representations of a sample. Given a sample, it should be pulled together by its positive samples. Conversely, it should be pushed apart with its negative samples as much as possible. From Fig. 3, wrongly predicted samples can be corrected by contrasting the change magnitude of different samples. Meanwhile, it can be seen that the correlation among different samples can be modeled by contrasting the Euclidean distance between their change representations, which indirectly optimizes the change magnitude of the samples. Thus, it is an effective alternative to the CMCL for metric-based models intuitively. The error $e$ in Eq. 4 can be replaced with

$$e_{k,k^+} = \mathrm{sqrt}(\sum_{i=1}^{C}(v_k^i - v_{k^+}^i)^2), \quad e_{k,k^-} = \mathrm{sqrt}(\sum_{i=1}^{C}(v_k^i - v_{k^-}^i)^2), \tag{5}$$

where $v_k^i$ represents the value in the $i$-th channel of the change representations of the sample $k$, and $C$ indicates the number of the feature channel.

Besides, we extend the CMCL so that it can be used for classification-based models. Without bells and whistles, the

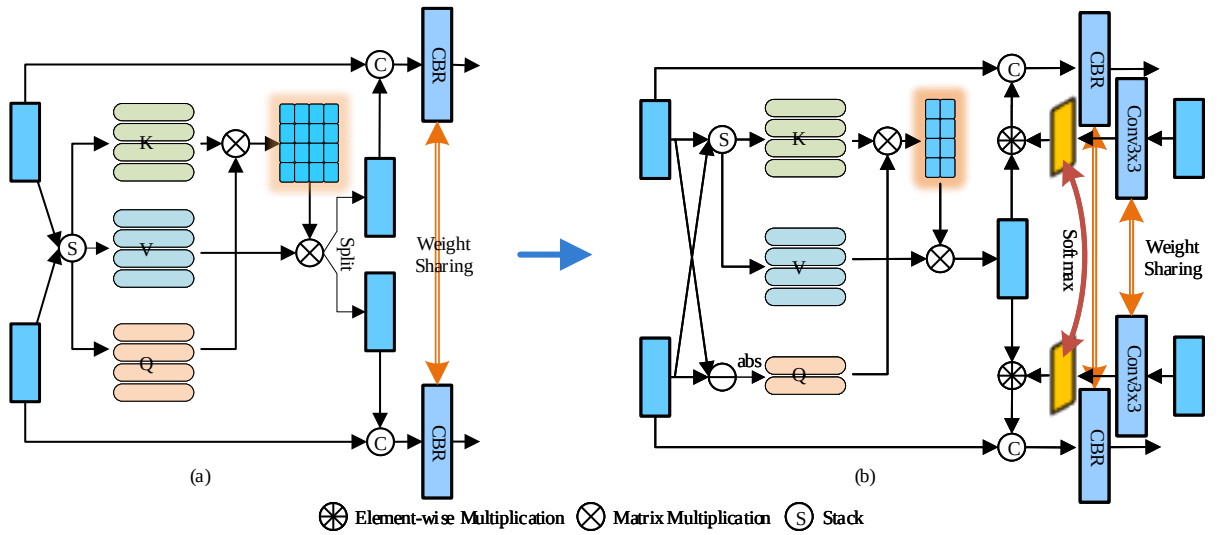



Fig. 4. The schematic diagram of the structure of the DGAM. (a) spatial-temporal attention mechanism (STAM), (b) difference-guided aggregation module (DGAM).

change confidence predicted by a classification-based model is regarded as the change magnitude of a sample. Here, the error $e$ in Eq. 4 can be formulated as

$$e_{k,k^+} = |p_k - p_{k^+}|, e_{k,k^-} = |p_k - p_{k^-}|, \tag{6}$$

where $p_k$ indicates the change confidence of the sample $k$. $p_k$ ranges between 0 and 1. Notably, the change magnitude of a sample is not known but is subjectively acquired according to the output of a model. Its reliability depends on the performance of the model. Since the CMCL is beneficial to reduce the change magnitude of the unchanged samples and increase that of the changed samples, the performance of a model can be enhanced by it.

### C. Difference-Guided Aggregation Module

A common structure of the STAM is presented in Fig. 4(a), where the query, the key, and the value are obtained from the concatenated bitemporal features in the spatial dimension. It can capture long-range dependency in bitemporal images and produce discriminative features with the guidance of semantic similarity. It does be beneficial to the improvement of the accuracy in change identification. However, it is prone to the confused aggregation of the bitemporal features with weak semantic discrimination, i.e., weak differences but different semantic categories, which can conversely reduce the consistency of the unchanged features or the discrepancy of changed features. To make up for its shortcoming, a difference-guided aggregation module (DGAM) is developed. We do not enhance but relax the semantic discrimination of the bitemporal representations.

The design of the DGAM is presented in Fig. 4(b). Let $F_{T1}$ and $F_{T2}$ represent the pre- and post-temporal features. Let $\mathrm{Q} \in \mathbb{R}^{C\times(h\times w)}$, $\mathrm{K} \in \mathbb{R}^{C\times(2\times h\times w)}$, and $\mathrm{V} \in \mathbb{R}^{C\times(2\times h\times w)}$ denote the query matrix, the key matrix, and the value matrix. The aggregation of the bitemporal features is formulated as

$$\mathrm{Q} = r(Conv1\mathrm{x}1(|F_{T1} - F_{T2}|)), \tag{7}$$

$$\mathrm{K} = r(Conv1\mathrm{x}1([F_{T1}, F_{T2}])), \tag{8}$$

$$\mathrm{V} = r(BR(Conv1\mathrm{x}1([F_{T1}, F_{T2}]))), \tag{9}$$

$$F_A = r(\mathrm{V} \otimes f(\mathrm{K}^{\mathrm{T}} \otimes \mathrm{Q})), \tag{10}$$

where $[\cdot,\cdot]$ indicates the stacked operation and $|\cdot|$ is to take the absolute value. $F_A \in \mathbb{R}^{C\times h\times w}$ denotes the aggregated features. $r(\cdot)$ represents the reshape operation and $f(\cdot)$ represents the softmax function. From Eqs. (7) to (10), it can be seen that DGAM learns the query from the absolute difference features of the bitemporal features. The query reflects the changed and unchanged embedding. Since the given labels represent the information of change, it is more discriminative than that learned from the bitemporal features. With its inducement and implicit supervision of given labels, the learned key strengthens the change relevance and relaxes the semantic distinction. Then the features relevant to the changed and unchanged representations can be aggregated to corresponding spatial positions in $F_A$ respectively. However, the $F_A$ can not be fused with the original bitemporal features directly. It can be inferred that the direct fusion is unfavorable to enhancing the discrepancy of bitemporal representations in changed regions. Consequently, an adaptive refined mechanism is introduced. Two attention maps are first obtained as follows

$$\begin{aligned} W_{T1}, W_{T2} =& f(Conv3\mathrm{x}3(F_{T1}), Conv3\mathrm{x}3(F_{T2})), \\ & s.t. W_{T1} + W_{T2} = \mathrm{I}, \end{aligned} \tag{11}$$

where $W_{T1}$ and $W_{T2}$ are the attention maps corresponding to the bitemporal inputs. I is the all-ones matrix. The sum of values of the $W_{T1}$ and $W_{T2}$ in the same position is equal to 1. Their values are adjusted dynamically according to the relation of the bitemporal features at the same position. Then, the $F_A$ is refined by the modulation of the attention maps, which is denoted as

$$\overline{F}_{T1} = F_A \circledast W_{T1}, \tag{12}$$

$$\overline{F}_{T2} = F_A \circledast W_{T2}, \tag{13}$$

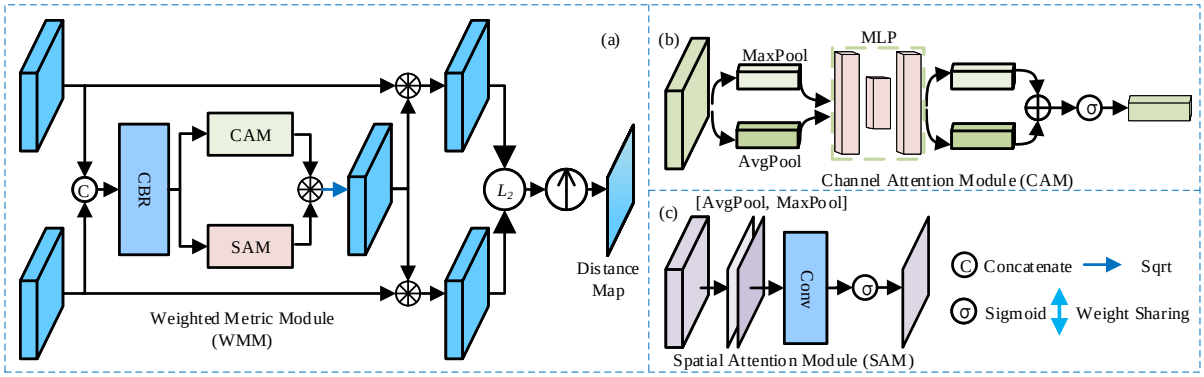



Fig. 5. The schematic diagram of the structure of the WMM. (a) weighted metric module (WMM), (b) channel attention module (CAM). (c) spatial attention module (SAM).

where $\overline{F}$ indicates the finely aggregated features. Finally, the $\overline{F}_{T1}$ and $\overline{F}_{T2}$ are concatenated and merged with the initial features $F_{T1}$ and $F_{T2}$ respectively. The operation of the feature merging (i.e., $CBR$) consists of one 3×3 convolutional layer and the $BR$. Here, the outputs of the DGAM not only capture the temporal-spatial correlation, but also enhance the differences of the bitemporal embedding in changed regions and the consistency of that in unchanged regions.

### D. Weighted Metric Module

The core of metric-based methods is to judge the changes according to the distance between the bitemporal features. On the one hand, the accuracy of the detection can be improved by extracting discriminative features. On the other hand, designing a reasonable metric strategy can achieve satisfactory results as well. Inspired by [26], we develop a weighted metric module (WMM). As presented in Fig. 5(a), it first learns multi-dimensional weights by employing the CAM and the SAM [35]. That can be formulated as follows

$$F = CBR(F_{pre}, F_{post}), \tag{14}$$

$$W_c = \mathrm{CAM}(F), \tag{15}$$

$$W_s = \mathrm{SAM}(F), \tag{16}$$

where $W_c$ represents the channel attention weights and $W_s$ indicates the spatial attention weights. $F_{pre}$ and $F_{post}$ represent the bitemporal features. From Eq. 14, the WMM learns the attention weights from their fused features $F$. Here, the WMM injects the state characteristics of the bitemporal pixels, i.e., change or non-change. Intuitively, it can well focus on the changed regions to suppress pseudo-changes. Then, the distance between the $F_{pre}$ and $F_{post}$ is flexibly computed with the variation of the index. Let $w_c^n \in W_c$ denotes the weight of the $n$-th channel and $w_s^{i,j} \in W_s$ denotes the weight of the $i$-th row and $j$-th column in the spatial dimension. The distance can be calculated as

$$d_{i,j} = \mathrm{sqrt}([(v_{1,i,j})^2 + \cdots (v_{n,i,j})^2 \cdots + (v_{N,i,j})^2]), \tag{17}$$

$$v_{n,i,j} = \max(\mathrm{sqrt}(w_s^{i,j} w_c^n), \epsilon)|q_{pre}^{n,i,j} - q_{post}^{n,i,j}|, \tag{18}$$

where $q^{n,i,j}$ represents the value of the bitemporal features and $v_{n,i,j}$ denotes the value of the change representations in the DGANet on the given index $(n, i, j)$ respectively. According to Eq. 17 and 18, the learned weights can dynamically remain or suppress the square error of the bitemporal features at every position to achieve adaptive distance computation, which is essentially the modulation to the difference representations of bitemporal embedding. Notably, to avoid resulting in the gradient explosion, a given parameter $\epsilon$ is used to constrain the minimum of the attention weight. Finally, the change map $X \in \mathcal{R}^{H \times W}$ is acquired by comparing the distance with a specific threshold, which is formulated as

$$x_{i,j} = \begin{cases} 1, & \text{if } d_{i,j} > \vartheta \\ 0, & \text{otherwise} \end{cases}, \tag{19}$$

where $\vartheta$ is the threshold and $x_{i,j}$ represents the predicted class in the position $(i, j)$.

## IV. Experiments

In this section, the evaluation metrics and datasets used to assess our approach are first given. Next, the quantitative and qualitative results are provided. In addition, the ablation study and discussion are performed. Finally, the hyperparameter analysis, the complexity analysis, and the limitation analysis are presented.

### A. Datasets and Evaluation Metrics

**Datasets:** We test the performance of the proposed method on three datasets, including the LEVIR-CD [25], the WHU-CD [53] and the SYSU-CD datasets [26]. The three change detection datasets are challenging and adopted widely. LEVIR-CD dataset is collected through the Google Earth platform. It is used to detect the construction and demolition of buildings, where various buildings are covered, e.g., warehouses and residences. 70% of the images in it are employed to train, 10% are to valid and 20% are to test. The images for training are randomly cropped into 256×256 patches during the training, while the images for validation or testing are clipped to 256×256 patches without overlapping.

WHU-CD dataset is an aerial image change detection dataset, the images in which own abundant details. The area that it covers occurred an earthquake in 2011, where many buildings are rebuilt or newly constructed in the following

years. In detail, it includes a pair of images of 32508×15354 size. We split it into two parts for training and testing respectively. The part for training is first cropped into 1024×1024 images without overlapping, and then 256×256 patches are selected randomly from them for training on every epoch. The part for testing is clipped directly to 256×256 images, where 2700 pairs of images are generated.

SYSU-CD dataset includes the aerial images acquired in Hong Kong from 2007 to 2014. It consists of multi-type changes across land and sea. Especially, there are many challenging scenarios, e.g., high-rise buildings and busy ports. It contains 20000 pairs of images of 256×256 size, which is generated by the data augmentation techniques including random flip and rotation. All images are divided into three groups on the basis of the ratio of 6:2:2, which are utilized for training, validation, and testing respectively.

**Evaluation Metrics:** Five common metrics are utilized to check the effectiveness of our model and reveal the differences among different methods, i.e., precision (P), recall (R), F1-score (F1), IoU, and overall accuracy (OA). These metrics except the OA refer to the predicted results of changed samples. Let $N_p$ and $N_{\overline{p}}$ represent the number of correctly and wrongly predicted changed samples, respectively, namely, true positives (TP) and false positives (FP). Let $N_f$ and $N_{\overline{f}}$ denote the number of correctly and wrongly predicted unchanged samples respectively, namely, true negatives (TN) and false negatives (FN). The above metrics can be demonstrated as follows:

$$P = N_p/(N_p + N_{\overline{p}}), \tag{20}$$

$$R = N_p/(N_p + N_{\overline{f}}), \tag{21}$$

$$IoU = N_p/(N_p + N_{\overline{f}} + N_{\overline{p}}), \tag{22}$$

$$F1 = \frac{2PR}{P+R}, \tag{23}$$

$$OA = \frac{N_p + N_f}{N_p + N_{\overline{p}} + N_f + N_{\overline{f}}}, \tag{24}$$

where F1 is the most critical indicator to objectively describe the effectiveness of the binary change detection algorithms.

### B. Implementation Details

The proposed model is realized founded on the PyTorch framework. The Adam optimizer is employed to optimize the DGANet, where the $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is 0.0001. We adjust the learning rate to half of its previous value every 40 epochs. There are a total of 200 epochs for training. During every forward computation, eight pairs of bitemporal images are fed into the model to acquire gradients and update the model weights, i.e., the batchsize is 8. We implement all models on one high-end NVIDIA GPU.

### C. Parameter Setting

The threshold $\vartheta$ in Eq. 19 on the two building change detection datasets is set to 2. Meanwhile, it is taken as 1 on the SYSU-CD dataset. The margin $\tau$ in the CMCL is 1 for a classification-based model. Differently, for metric-based models explored in this work, it is 2, and too when the Eq. 5 is used to contrast different samples. The balanced coefficient $\gamma$ of the CMCL is set to 0.1. Besides, the $\epsilon$ in Eq. 18 is 1e-12, which avoids gradient explosion possibly caused by the attention weights around zero in WMM and is enough small not to interfere with the role of WMM. Following previous work [26], [52], the margin $\Gamma$ in the BCL is set to 2. In following experiments without special instructions, these parameter settings are adopted by default.

### D. Comparison with Advanced Methods

There are nine advanced methods introduced to verify the superiority of our method. They are briefly illustrated in the following:

1) FC-EF [16]: A FCN architecture that detects the change from the concatenated bitemporal images in the band dimension.
2) FC-Siam-diff [16]: A Siamese FCN model that acquires change-related information from the absolute difference features of the multi-level bitemporal features.
3) FC-Siam-conc [16]: Different from the FC-Siam-diff, it extracts the information from the concatenated bitemporal features to generate the change map.
4) STANet [25]: A algorithm based on metric-learning that uses the multi-scale bitemporal attention module to improve the accuracy of change identification.
5) DSAMNet [26]: A metric-based method that introduces two independent convolutional block attention modules to refine the bitemporal features respectively at the decision stage. Besides, it adopts the deep supervised strategy to learn discriminative features.
6) SNUNet [54]: Dense skip connections are employed to maintain changed object position information in the SNUNet. Moreover, an integrated channel attention mechanism is designed to modulate the multi-level features.
7) EGRCNN [55]: It is dedicated to building change detection, where the edge prior is employed by introducing the edge prediction task to maintain the integrity of the building structure.
8) ChangeFormer [56]: It uses the advanced Vision Transformer (ViT) as the encoder, which naturally captures the long-range dependency of the bitemporal images. Similar to the FC-Siam-conc, the changed features are generated by concatenating the bitemporal features.
9) BIT [29]: A hybrid architecture method that uses the CNN as the encoder and the Transformer as the decoder. In particular, it designs a bitemporal Transformer module to exploit the temporal-spatial correlation.

**Results on the LEVIR-CD dataset:** Fig. 6 presents the visual comparison of different change detection methods. From the given three pairs of images, our method achieves the best prediction results overall. The changed building presented in the first row covers a large area with an irregular distribution. It can be seen that the color of the building is close to that of the land surface, where the methods introducing the edge knowledge or the spatial attention can effectively increase the

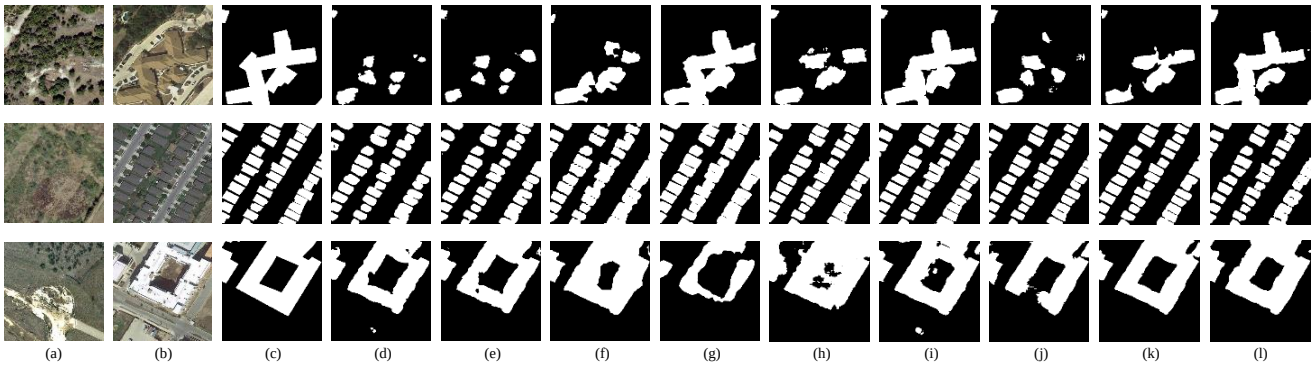



Fig. 6. The visual comparison of different methods on the LEVIR-CD dataset. (a) Pre-temporal image, (b) Post-temporal image, (c) Ground truth, (d) FC-Siam-diff, (e) FC-Siam-conc, (f) STANet, (g) DSAMNet, (h) SNUNet, (i) EGRCNN, (j) ChangeFormer, (k) BIT, (l) Ours.

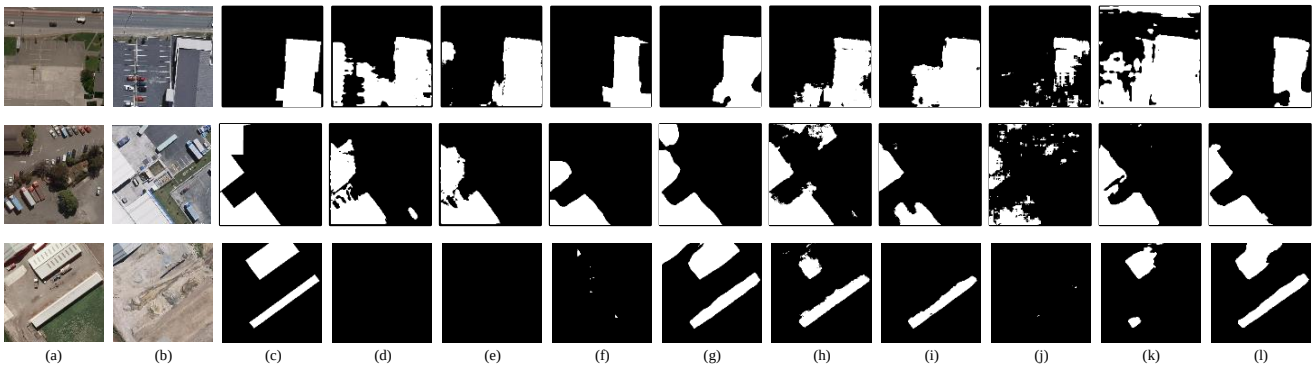



Fig. 7. The visual comparison of different methods on the WHU-CD dataset. (a) Pre-temporal image, (b) Post-temporal image, (c) Ground truth, (d) FC-Siam-diff, (e) FC-Siam-conc, (f) STANet, (g) DSAMNet, (h) SNUNet, (i) EGRCNN, (j) ChangeFormer, (k) BIT, (l) Ours.

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

| Method | Evaluation Metrics | | | | |
|---|---|---|---|---|---|
| | P | R | F1 | IoU | OA |
| FC-EF [16] | 86.09 | 83.50 | 84.77 | 73.57 | 98.47 |
| FC-Siam-diff [16] | 90.05 | 83.48 | 86.64 | 76.43 | 98.69 |
| FC-Siam-conc [16] | 90.46 | 83.84 | 87.02 | 77.03 | 98.73 |
| STANet [25] | 85.01 | **91.38** | 88.07 | 78.69 | 98.74 |
| DSAMNet [26] | 81.28 | 88.68 | 84.81 | 73.63 | 98.38 |
| SNUNet [54] | 90.69 | 88.93 | 89.80 | 81.49 | 98.97 |
| EGRCNN [55] | 88.58 | 91.13 | 89.84 | 81.55 | 98.95 |
| ChangeFormer [56] | 91.03 | 87.77 | 89.37 | 80.78 | 98.94 |
| BIT [29] | 91.61 | 88.74 | 90.15 | 82.07 | 99.01 |
| DGANet | **92.03** | 88.56 | **90.26** | **82.25** | **99.03** |

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

| Method | Evaluation Metrics | | | | |
|---|---|---|---|---|---|
| | P | R | F1 | IoU | OA |
| FC-EF [16] | 82.84 | 78.02 | 80.36 | 67.16 | 98.62 |
| FC-Siam-diff [16] | 73.27 | 82.86 | 77.77 | 63.62 | 98.29 |
| FC-Siam-conc [16] | 65.41 | 86.32 | 74.42 | 59.27 | 97.85 |
| STANet [25] | 88.59 | 85.18 | 86.86 | 76.76 | 99.07 |
| DSAMNet [26] | 71.22 | **92.28** | 80.40 | 67.22 | 98.37 |
| SNUNet [54] | 83.95 | 88.95 | 86.38 | 76.02 | 98.99 |
| EGRCNN [55] | 90.92 | 89.41 | 90.16 | 82.08 | 99.29 |
| ChangeFormer [56] | 88.22 | 79.86 | 83.83 | 72.16 | 98.89 |
| BIT [29] | 78.33 | 89.21 | 83.42 | 71.55 | 98.72 |
| DGANet | **95.59** | 86.59 | **90.87** | **83.27** | **99.37** |

discrimination of the bitemporal features. The second row shows dense buildings. The changed buildings generated by the STANet and DSAMNet almost connect together. Compared with them, our method alleviates the problem to some extent, which can be attributed to the effect of the WMM and the CMCL. The last example presents one changed object with a hole and the interference of the shadow. Some methods (e.g., DSAMNet and ChangeFormer) are affected by shadows, and the changed samples adjacent to the shadow are missed. In general, our method keeps the compactness of the changed objects and mitigates the negative impact of external factors including the angle of the illumination.

Table I gives the quantitative results. Our method attains the highest F1 of 90.26%, which is slightly superior to the F1 of the BIT. Besides, the highest precision of 92.03% is obtained with an acceptable recall of 88.56%. From the visual comparison and the evaluation metrics, the performance of DGANet on the LEVIR-CD dataset is relatively competitive.

**Results on the WHU-CD dataset:** The visual comparison of different change detection methods is shown in Fig. 7. The image pair in the first row shows the color pseudo-change of the land surface, which is effectively filtered by metric-based methods containing the STANet, DSAMNet, and ours. The BIT even generates more false detection. It can be speculated that BIT adversely enhances the differences of the representations in unchanged regions, due to similar

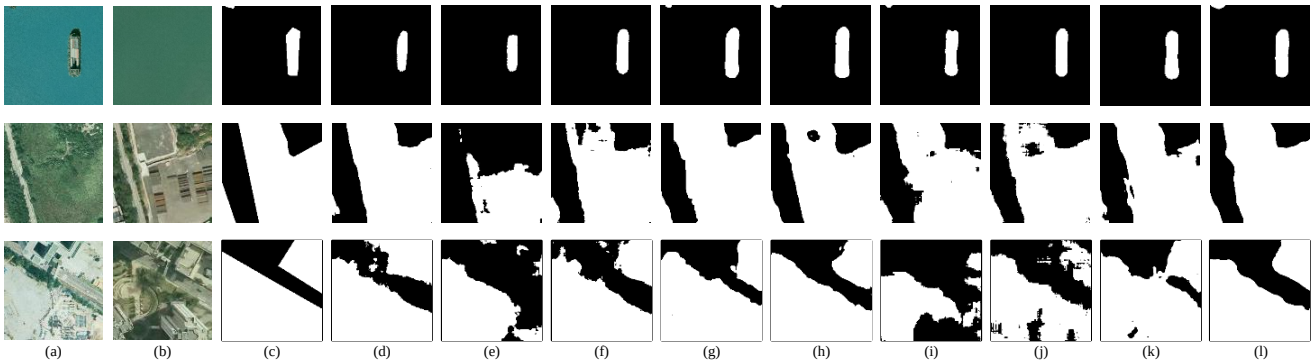



Fig. 8. The visual comparison of different methods on the SYSU-CD dataset. (a) Pre-temporal image, (b) Post-temporal image, (c) Ground truth, (d) FC-EF, (e) FC-Siam-diff, (f) FC-Siam-conc, (g) STANet, (h) DSAMNet, (i) SNUNet, (j) ChangeFormer, (k) BIT, (l) Ours.

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

| Method | Evaluation Metrics | | | | |
|---|---|---|---|---|---|
| | P | R | F1 | IoU | OA |
| FC-EF [16] | 79.26 | 75.50 | 77.34 | 63.05 | 89.56 |
| FC-Siam-diff [16] | **90.30** | 53.48 | 67.17 | 50.57 | 87.67 |
| FC-Siam-conc [16] | 78.40 | 75.82 | 77.09 | 62.72 | 89.37 |
| STANet [25] | 74.97 | 80.53 | 77.65 | 63.48 | 89.06 |
| DSAMNet [26] | 72.22 | **83.07** | 77.27 | 62.96 | 88.47 |
| SNUNet [54] | 82.71 | 71.96 | 76.96 | 62.55 | 89.84 |
| EGRCNN [55] | N/A | N/A | N/A | N/A | N/A |
| ChangeFormer [56] | 81.30 | 74.65 | 77.83 | 63.71 | 89.97 |
| BIT [29] | 80.94 | 75.41 | 78.08 | 64.04 | 90.01 |
| DGANet | 78.86 | 80.74 | **79.79** | **66.37** | **90.35** |

spectral characteristics between the ground and the building in the post-temporal image. The DGANet fails to identify the upper-left building in the second example. Instead, the DSAMNet predicts it successfully, which can demonstrate that it is sensitive to subtle change. Our method captures another changed building completely, which is omitted by the ChangeFormer. The third row presents two disappeared buildings. Here, our method recalls all buildings with high precision. Note that the DSAMNet produces false detection, where the building with the slight difference is judged as the changed object. Besides, the SNUNet fails to retain the integrity of the changed buildings.

The quantitative results are reported in Table II. It can be seen that there is a clear imbalance between the precision and recall of chosen methods except for the EGRCNN. In detail, some methods obtain a higher recall but with low precision. Instead, our method owns the lower recall than the precision. Due to the limitation of the strict threshold, there are relatively more omissions in our model. The positive side is that it is beneficial to suppress the change irrelevant to the building. In detail, the highest precision of 95.59% is achieved, and the F1 of 90.87% is the highest. The EGRCNN obtains the second-best F1 of 90.16%. Its precision and recall are close, which illustrates that edge knowledge is helpful to improve recall while keeping good precision. In general, our model achieves comparable performance.

**Results on the SYSU-CD dataset:** The visual results of different change detection methods are presented in Fig. 8. The first pair of images shows the change of one ship and one tiny changed region. Here, due to the clean sea presented by the post-temporal image, the advantage of the dense skip connection to reserve the shallow information is highlighted. Meanwhile, the STANet, DSAMNet, SNUNet, and ours locate the latter successfully. The second pair of images present multiple types of land cover change, including buildings, roads, and vegetation. The prediction of the FC-Siam-diff is worst. The images in the third row show a challenging scenario, where the two images have different views and changed high-rise buildings are required to extract. The visual comparison is close to that of the second pair of images. Note that the SNUNet, ChangeFormer, and BIT miss some changed samples, which possibly is owing to the influence of the shadow. Our model better ensures the predictive integrity of the defined regions of change overall.

The quantitative results are recorded in Table III. In particular, The quantitative results of the FC-Siam-diff match with the visual results. Our method produces the highest F1 of 79.79%, and the precision and the recall are acceptable. It can be noted that the F1 of multiple methods is far lower than that on another two datasets, which can be attributed to puzzling changed objects and challenging scenes.

In summary, DGANet provides satisfactory performance compared with the selected advanced methods. DGANet can effectively reduce the interference of external factors, such as illumination, view, and season variation. Moreover, DGANet can better keep the completeness of the changed objects.

### *E. Ablation Study and Discussion*

The ablation study is performed on the LEVIR-CD dataset. All possible combinations of the three improvements are investigated to confirm their effectiveness. In detail, Tabel. IV reports the experimental results, where eight experiments are implemented. No. 1 indicates the baseline model. From No. 2 to 8, the effectiveness of the three modules can be illustrated by the increase of the F1. Notably, compared with the baseline model, the recall of the models only introducing the DGAM or the WMM decreases, and their precision increases. That reveals that the DGAM and the WMM can suppress the pseudo-changes well. However, more TPs are missed. An

TABLE IV
ABLATION STUDIES AND COMPARISON ABOUT THE MODULES (%). THE BEST RESULTS ARE IN BOLD.

| No. | Module | | | Evaluation Metrics | | | | |
|---|---|---|---|---|---|---|---|---|
| | CMCL | DGAM | WMM | P | R | F1 | IoU | OA |
| 1 | ✗ | ✗ | ✗ | 89.05 | 87.27 | 88.15 | 78.81 | 98.80 |
| 2 | ✗ | ✓ | ✗ | 91.19 | 86.32 | 88.69 | 79.68 | 98.89 |
| 3 | ✗ | ✗ | ✓ | 92.29 | 85.40 | 88.71 | 79.71 | 98.89 |
| 4 | ✗ | ✓ | ✓ | 92.39 | 86.47 | 89.33 | 80.72 | 98.95 |
| 5 | ✓ | ✗ | ✗ | 87.26 | **89.57** | 88.40 | 79.22 | 98.80 |
| 6 | ✓ | ✓ | ✗ | 92.08 | 86.68 | 89.30 | 80.67 | 98.94 |
| 7 | ✓ | ✗ | ✓ | **92.97** | 86.96 | 89.86 | 81.59 | 99.00 |
| 8 | ✓ | ✓ | ✓ | 92.03 | 88.56 | **90.26** | **82.25** | **99.03** |

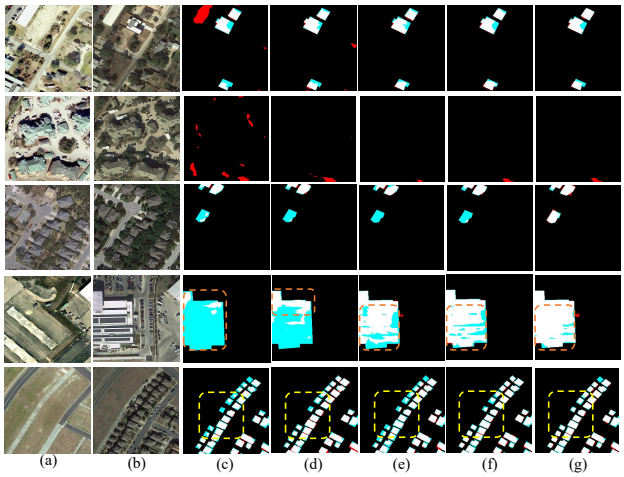



Fig. 9. The visualization of ablation studies. (a) Pre-temporal image. (b) Post-temporal image. (c) Baseline. (d) Baseline+DGAM. (e) Baseline+WMM. (f) Baseline+DGAM+WMM. (g) Baseline+DGAM+WMM+CMCL. The white, black, red, and blue refer to the TPs, TNs, FPs, and FNs respectively.

interpretable reason is that DGAM and WMM reduce the distinction of bitemporal representations in the changed regions with weak differences. The above can be further demonstrated according to the first three examples in Fig. 9. Actually, DGAM requires well-differences representations to guide the non-local aggregation between bitemporal embedding. From Fig.9, on the basis of the baseline model, the model injecting DGAM can effectively improve the accuracy of change identification in the regions with similar differences between bitemporal images. Similarly, according to the comparison between No. 2 and No. 4, and the comparison between No.6 and No. 8, with the help of enhanced bitemporal representations, WMM can better facilitate the detection of changed regions. In addition, compared with No. 4, the introduction of CMCL in No. 8 significantly enhances the performance of the model, and Fig. 9 shows the prediction of changed regions is more complete. Overall, the three modules are helpful to improve the accuracy of change identification.

To further illustrate the effectiveness of the DGAM, the compared experiments between the STAM presented in Fig. 4(a) and the DGAM are executed. The performance of the STAM is explored by replacing the DGAM with it in the DGANet. Table. V reports the quantitative results. The results on the LEVIR-CD and WHU-CD datasets verify the two modules are helpful to improve the performance of the model.

TABLE V
THE COMPARISON BETWEEN THE STAM AND THE DGAM (%).

| Dataset | Module | | Evaluation Metrics | |
|---|---|---|---|---|
| | STAM | DGAM | F1 | IoU |
| LEVIR-CD | – | – | 89.86 | 81.59 |
| | ✓ | – | 90.15 | 82.06 |
| | – | ✓ | 90.26 | 82.25 |
| WHU-CD | – | – | 89.95 | 81.74 |
| | ✓ | – | 90.20 | 82.14 |
| | – | ✓ | 90.87 | 83.27 |
| SYSU-CD | – | – | 78.71 | 64.90 |
| | ✓ | – | 78.69 | 64.86 |
| | – | ✓ | 79.79 | 66.37 |

TABLE VI
THE EXPLORATION ON THE GENERALITY OF THE CMCL (%).

| Method | Euclidean Distance | Change Magnitude | Evaluation Metrics | |
|---|---|---|---|---|
| | | | F1 | IoU |
| DGANet | – | – | 89.33 | 80.72 |
| | ✓ | – | 90.06 | 81.91 |
| | – | ✓ | 90.26 | 82.25 |
| DSAMNet | – | – | 84.81 | 73.63 |
| | ✓ | – | 86.10 | 75.59 |
| | – | ✓ | 85.58 | 74.80 |
| FC-EF | – | – | 84.77 | 73.57 |
| | – | ✓ | 85.39 | 74.51 |
| FC-Siam-diff | – | – | 86.64 | 76.43 |
| | – | ✓ | 86.90 | 76.84 |
| FC-Siam-conc | – | – | 87.02 | 77.03 |
| | – | ✓ | 87.61 | 77.96 |

Actually, the given labels of the two datasets possess explicit semantic information, i.e., building. Therefore, the STAM well presents its function. However, semantic classes in changed regions are not indicated by provided labels on the SYSU-CD dataset, thus it can be inferred that the semantic discrimination of representations learned by the model is weak. Meanwhile, the STAM fails to bring performance gain. Conversely, DGAM relaxes the semantic discrimination of bitemporal representations and emphasizes discriminative difference representations. Its effectiveness is demonstrated by the obvious performance improvement on the SYSU-CD dataset and some visualized results shown in Fig. 10.

Furthermore, we investigate the generality of the CMCL. Tabel. VI reports the relevant results on the LEVIR-CD dataset. The DGANet and the DSAMNet are metric-based models. The FC-EF, the FC-Siam-diff, and the FC-Siam-conc are classification-based models. First, for the two metric-based models, we demonstrate the effectiveness of employing the Euclidean distance between the change representations to model the dependency among the bitemporal pixels. It improves the performance of the DSAMNet and that of the DGANet. Second, it can be noted that the CMCL can improve the performance of both the two metric-based models and the three classification-based models simultaneously, which indicates our extension to the CMCL is effective. In brief, modeling the correlation among bitemporal pixes is helpful to enhance the representation quality of a model, thus the accuracy of change detection can be improved.

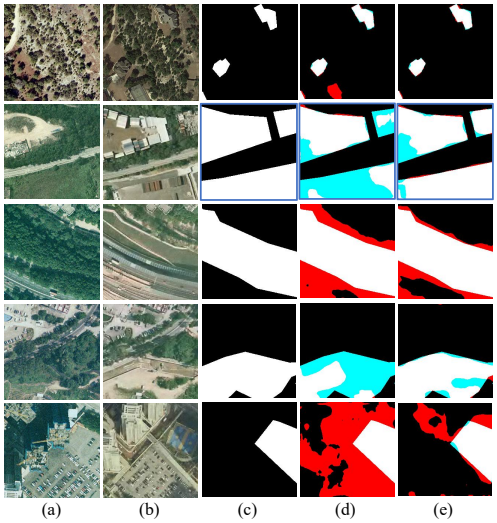



Fig. 10. The visual comparison between the DGAM and the STAM. (a) Pre-temporal image. (b) Post-temporal image. (c) Ground Truth. (d) Baseline+WMM+CMCL+STAM. (e) Baseline+WMM+CMCL+DGAM. The white, black, red, and blue refer to the TPs, TNs, FPs, and FNs respectively.

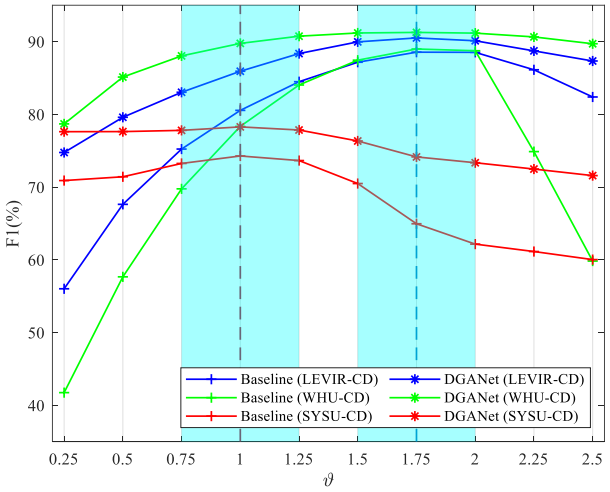



Fig. 11. The best F1-score of the baseline model and DGANet under different thresholds on the test (WHU-CD) or validation data of the three datasets.

### F. Hyperparameter Analysis

In this section, we focus on exploring the effect of the $\vartheta$, $\gamma$, and $\tau$ by the variable-control method. Moreover, the rationality of their selection in this work is illustrated.

**The threshold $\vartheta$:** According to Eq. 19, the $\vartheta$ is used to judge if the bitemporal pixels are changed. It usually is 0.5 by default in classification-based methods. Its selection in metric-based models is different from that. Here, we investigate the influence of its variation on the accuracy of the proposed model. The results are presented in Fig. 11, where it can be seen that the $\vartheta$ plays a key role in accurately identifying changes. Firstly, the determination of $\vartheta$ is closely relevant to the $\Gamma$ in the BCL (Eq. 1). The $\Gamma$ is set to 2. The BCL optimizes the bitemporal representations in changed regions only if the distance between them is less than 2. Therefore, if $\vartheta$ is more than 2, many changed pixels cannot be detected. Secondly, from Fig. 11, a threshold close to the $\Gamma$ can make the baseline model and the DGANet achieve the highest accuracy on the two building change detection datasets. Meanwhile, on the SYSU-CD dataset, the highest accuracy can be achieved when the threshold is close to the medium of the $\Gamma$. Essentially, for the specific-type and multi-type change detection, the distributions of the distance between the bitemporal embedding learned in interested changed regions are different. Thus, the selection of the threshold should take the change types indicated on the dataset into consideration. In this work, to well reserve interested changed pixels and exclude unchanged or irrelevant changed pixels, the $\vartheta$ on the two building change datasets is 2. It is 1 on the SYSU-CD dataset. They reveal the performance of the proposed model objectively.

TABLE VII
THE ANALYSIS ON THE TUNED COEFFICIENT $\gamma$ OF THE CMCL.

| $\gamma$ | DGANet ($\tau$=2) | | FC-EF ($\tau$=1) | |
|---|---|---|---|---|
| | F1 | IoU | F1 | IoU |
| – | 89.33 | 80.72 | 84.77 | 73.57 |
| 0.05 | 89.84 | 81.55 | 85.12↑ | 74.09↑ |
| 0.1 | 90.26 | 82.25 | **85.39**↑ | **74.51**↑ |
| 0.2 | 89.92 | 81.69 | 84.70 | 73.46 |
| 0.3 | 89.93 | 81.70 | 84.37 | 72.96 |
| 0.5 | **90.48** | **82.62** | 83.79 | 72.10 |
| 0.7 | 89.96 | 81.76 | 83.79 | 72.10 |
| 10.0 | 90.00 | 81.81 | 82.84 | 70.70 |

TABLE VIII
THE ANALYSIS ON THE MARGIN $\tau$ IN CMCL. ($\gamma$=0.1)

| $\tau$ | DGANet | | FC-EF | |
|---|---|---|---|---|
| | F1 | IoU | F1 | IoU |
| – | 89.33 | 80.72 | 84.77 | 73.57 |
| 0.25 | 89.01 | 80.20 | 84.65 | 73.38 |
| 0.5 | 89.45 | 80.91 | 85.30 | 74.37 |
| 1 | 89.27 | 80.63 | **85.39** | **74.51** |
| 2 | **90.26** | **82.25** | 85.27 | 74.32 |
| 4 | 90.13 | 82.03 | 85.14 | 74.13 |

**The tuned coefficient $\gamma$:** The $\gamma$ is used to regulate the influence of the CMCL on the optimization of the model. Here, how the variation of the $\gamma$ affects the performance of the model is explored. Table. VII reports the experimental results of the DGANet and the FC-EF on the LEVIR-CD dataset. It can be noted that the accuracy of the FC-EF is improved when the $\gamma$ does not exceed 0.1. As the $\gamma$ is more than 0.2, its accuracy decreases. In particular, an extreme situation (i.e., $\gamma$=10) is given, where the performance of the FC-EF degrades obviously. Although the performance of the DGANet can be improved under different $\gamma$, there is a gap in the performance gain. Therefore, if the effect of the CMCL cannot be balanced well, there is an increased risk of performance degradation. In fact, according to Eqs. 3 and 4, the CMCL is helpful to push the changed samples and the unchanged samples apart. However, it cannot distinguish their categories. The update of the model weights to make it decline possibly makes the pixel-wise supervised loss increase instead, which indicates more samples are predicted wrongly. In this work, the tuned coefficient $\gamma$ is set to 0.1. The experimental results demonstrate it is a reasonable setting and can make the CMCL bring a positive effect for different models.

**The marign $\tau$:** Table. VIII reports the influence of the

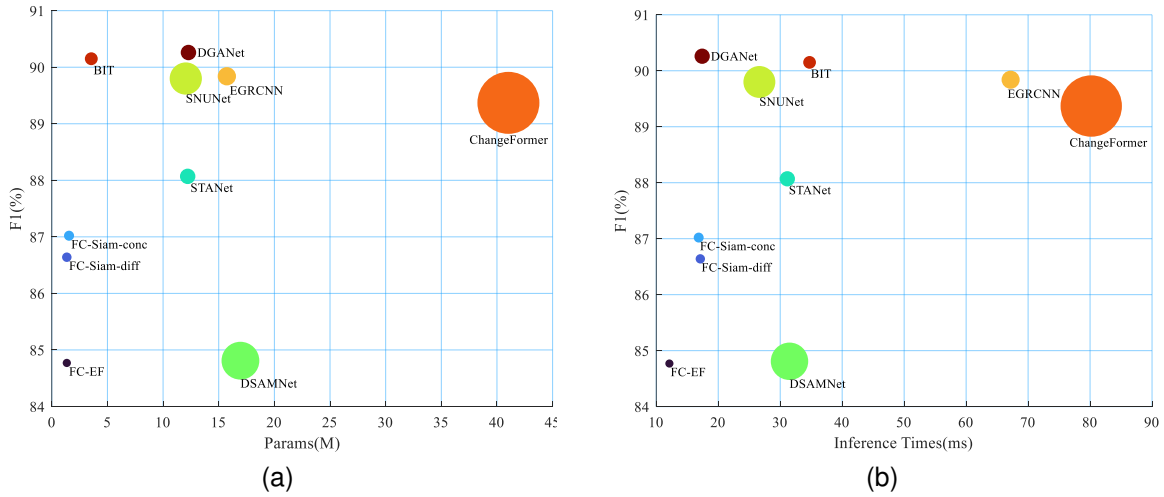



Fig. 12. The comparison of the Params, the inference time, FLOPs, and F1 for different models. The area of the circle indicates the size of FLOPs.

margin $\tau$ in CMCL on the performance of the two typical models, the proposed metric-based model DGANet and the classification-based model FC-EF. The experiments are implemented on the LEVIR-CD dataset. It can be seen that if the $\tau$ is less than 2, the CMCL hardly brings performance improvement for the DGANet. For the FC-EF, when the $\tau$ is not less than 0.5, the CMCL improves its accuracy. Actually, the margin $\tau$ determines whether the CMCL can well separate the samples with different categories. According to the value of the $\Gamma$ in BCL used for the DGANet, it is hoped that the distance between the bitemporal features is close to 0 in the unchanged samples and larger than 2 in the changed samples. Meanwhile, for the classification-based model, it is wanted that the change confidence is close to 0 in unchanged regions and 1 in changed regions. Therefore, the margin $\tau$ in CMCL is set to 1 for the classification-based model. It is set to 2 when the CMCL is employed to optimize the DGANet. They are intuitive settings that are enough to push apart the samples with different categories. Besides, from Table. VI, the rationality of the setting on the $\tau$ is further demonstrated, where the parameter settings of the CMCL used for the DSAMNet are the same as those in the DGANet. Notably, they are also optimized by the identical BCL. In summary, the CMCL with reasonable parameter settings can well enhance the performance of different types of models.

### G. Complexity Analysis

To comprehensively evaluate the performance of our approach, the number of model parameters (Params), the floating-point operations (FLOPs), the inference time consumption, and the F1 are counted on the LEVIR-CD dataset. Table. IX records the specific results of the comparison networks and our model. The inference time refers to the average time consumption of all images used for testing on the LEVRI-CD dataset. Besides, Fig. 12 gives an intuitive comparison from the above four perspectives. The STANet has close Params and FLOPs with the DGANet, but its F1 is far lower. Besides, although the ChangeFormer achieves a good F1, it brings a large computational burden because of the high complexity of the Transformer architecture. From Fig. 12, the accuracy of our model is slightly better than that of the BIT, where the latter has fewer Params and lower FLOPs. Nevertheless, the inference speed of the DGANet is faster, which can be attributed to the well-designed operations, such as the difference-guided fusion of the bitemporal features at the highest level. As a whole, compared with these advanced methods, our model achieves the highest F1 with competitive complexity.

TABLE IX
THE ANALYSIS OF THE COMPLEXITY OF DIFFERENT NETWORKS ON THE LEVIR-CD DATASET. THE BEST RESULTS ARE IN BOLD.

| Method | Params(M) | FLOPs(G) | Inference Time(ms) | F1(%) |
|---|---|---|---|---|
| FC-EF [16] | **1.35** | **3.58** | **12.12** | 84.77 |
| FC-Siam-diff [16] | **1.35** | 4.73 | 17.11 | 86.64 |
| FC-Siam-conc [16] | 1.55 | 5.33 | 16.85 | 87.02 |
| STANet [25] | 12.21 | 12.56 | 31.15 | 88.07 |
| DSAMNet [26] | 16.95 | 75.39 | 31.50 | 84.81 |
| SNUNet [54] | 12.04 | 54.83 | 26.64 | 89.80 |
| EGRCNN [55] | 15.73 | 17.64 | 67.18 | 89.84 |
| ChangeFormer [56] | 41.03 | 202.79 | 80.19 | 89.37 |
| BIT [29] | 3.55 | 8.75 | 34.74 | 90.15 |
| DGANet | 12.28 | 12.56 | 17.41 | **90.26** |

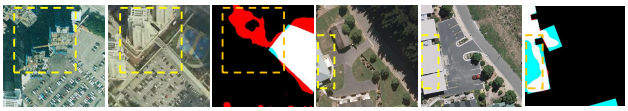


Fig. 13. The visualization of some failed cases. The white, black, red, and blue refer to the TPs, TNs, FPs, and FNs respectively.

### H. Limitation Analysis

Some failed cases are given in Fig. 13, which are used to illustrate the limitation of our method. In the first example, there is a pair of off-nadir images taken from different views. Since the pixels of the high-wise buildings in it are not well aligned, false alarms are triggered. Besides, the change between the building instances is hard to recognize, i.e., from one

building to another building. Essentially, our method identifies changes by measuring the difference between representations of the same position in the bitemporal images. Therefore, it requires well-registered near-nadir images to detect changes. Meanwhile, it is uneasy for it to identify the change between the specific-type object instances.

## V. Conclusion

In this work, a novel metric-based model DGANet is developed for remote sensing image change detection. It injects two newly developed modules: the DGAM and the WMM. DGAM achieves the discriminative fusion of the bitemporal features by the interaction between the difference features and the bitemporal features. WMM dynamically measures the distance between the bitemporal features by adaptive feature attention in multiple dimensions, which effectively suppresses the pseudo-changes. Moreover, a change magnitude contrastive loss is introduced and extended, which explores and employs the relationship of bitemporal pixels within multiple bitemporal images. It further improves the representation discrimination of the model. The extensive comparison experiments and ablation studies demonstrate the effectiveness of our method. However, it is challenging for it to identify changes in off-nadir images and distinguish different object instances. We will explore solutions in future work.

## References


[1] H. Luo, C. Liu, C. Wu, and X. Guo, "Urban change detection based on dempster–shafer theory for multitemporal very high-resolution imagery," *Remote Sensing*, vol. 10, no. 7, p. 980, 2018.

[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. Howarth and G. Wickware, "Procedures for change detection using landsat digital data," *International Journal of Remote Sensing*, vol. 2, no. 3, pp. 277–291, 1981.

[4] 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.

[5] 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.

[6] 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.

[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] D. Seo, Y. Kim, Y. Eo, M. Lee, and W. Park, "Fusion of sar and multispectral images using random forest regression for change detection," *ISPRS International Journal of Geo-Information*, vol. 7, no. 10, p. 401, 2018.

[9] 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.

[10] X. Zhang, P. Xiao, X. Feng, and M. Yuan, "Separate segmentation of multi-temporal high-resolution remote sensing images for object-based change detection in urban area," *Remote Sensing of Environment*, vol. 201, pp. 243–255, 2017.

[11] Y. Tang, X. Huang, and L. Zhang, "Fault-tolerant building change detection from urban high-resolution remote sensing imagery," *IEEE geoscience and remote sensing letters*, vol. 10, no. 5, pp. 1060–1064, 2013.

[12] T. Lei, Y. Zhang, Z. Lv, S. Li, S. Liu, and A. Nandi, "Landslide inventory mapping from bitemporal images using deep convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 6, pp. 982–986, 2019.

[13] Q. Wang, Z. Yuan, Q. Du, and X. Li, "GETNET: A general end-to-end 2-D CNN framework for hyperspectral image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 3–13, 2019.

[14] G. Liu, Y. Yuan, Y. Zhang, Y. Dong, and X. Li, "Style transformation-based spatial–spectral feature learning for unsupervised change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[15] T. Lei, J. Wang, H. Ning, X. Wang, D. Xue, Q. Wang, and A. Nandi, "Difference enhancement and spatial–spectral nonlocal network for change detection in vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[16] R. Daudt, B. Le, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proc. IEEE International Conference on Image Processing*, 2018, pp. 4063–4067.

[17] M. Papadomanolaki, S. Verma, M. Vakalopoulou, S. Gupta, and K. Karantzalos, "Detecting urban changes with recurrent neural networks from multitemporal sentinel-2 data," in *Proc. IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019, pp. 214–217.

[18] 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.

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

[20] J. Liu, M. Gong, K. Qin, and P. Zhang, "A deep convolutional coupling network for change detection based on heterogeneous optical and radar images," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 3, pp. 545–559, 2016.

[21] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep siamese convolutional network for optical aerial images," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1845–1849, 2017.

[22] M. Liu, Q. Shi, A. Marinoni, D. He, X. Liu, and L. Zhang, "Super-resolution-based change detection network with stacked attention module for images with different resolutions," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2022.

[23] Z. Cai, Z. Jiang, and Y. Yuan, "Task-related self-supervised learning for remote sensing image change detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 1535–1539.

[24] M. Zhang, G. Xu, K. Chen, M. Yan, and X. Sun, "Triplet-based semantic relation learning for aerial remote sensing image change detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 2, pp. 266–270, 2018.

[25] 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.

[26] 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.

[27] Y. Zhou, F. Wang, J. Zhao, R. Yao, S. Chen, and H. Ma, "Spatial-temporal based multi-head self-attention for remote sensing image change detection," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2022.

[28] 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.

[29] 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.

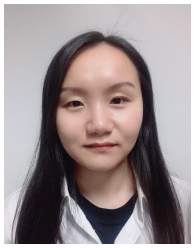
[30] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 076–10 085.

[31] Q. Li, M. Gong, Y. Yuan, and Q. Wang, "Symmetrical feature propagation network for hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[32] Q. Li, Y. Yuan, X. Jia, and Q. Wang, "Dual-stage approach toward hyperspectral image super-resolution," *IEEE Transactions on Image Processing*, vol. 31, pp. 7252–7263, 2022.
[33] X. Zhu, D. Cheng, Z. Zhang, S. Lin, and J. Dai, "An empirical study of spatial attention mechanisms in deep networks," in *Proc. IEEE international conference on computer vision*, 2019, pp. 6688–6697.
[34] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.
[35] S. Woo, J. Park, J. Lee, and I. Kweon, "Cbam: Convolutional block attention module," in *Proc. European conference on computer vision*, 2018, pp. 3–19.
[36] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
[38] J. Huang, Q. Shen, M. Wang, and M. Yang, "Multiple attention siamese network for high-resolution image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021.
[39] Z. Lv, F. Wang, G. Cui, J. Benediktsson, T. Lei, and W. Sun, "Spatial–spectral attention network guided with change magnitude image for land cover change detection using remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.
[40] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, and G. Liu, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 183–200, 2020.
[41] Z. Wu, Y. Xiong, S. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE conference on computer vision and pattern recognition*, 2018, pp. 3733–3742.
[42] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.
[43] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.
[44] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9912–9924, 2020.
[45] S. Saha, P. Ebel, and X. X. Zhu, "Self-supervised multisensor change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–10, 2022.
[46] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 546–12 558, 2020.
[47] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, "Dense contrastive learning for self-supervised visual pre-training," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3024–3033.
[48] Z. Xie, Y. Lin, Z. Zhang, Y. Cao, S. Lin, and H. Hu, "Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 684–16 693.
[49] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 661–18 673, 2020.
[50] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. Van Gool, "Exploring cross-image pixel contrast for semantic segmentation," in *Proc. IEEE International Conference on Computer Vision*, 2021.
[51] T. Zhou, M. Zhang, F. Zhao, and J. Li, "Regional semantic contrast and aggregation for weakly supervised semantic segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4299–4309.
[52] 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.
[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] 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.
[55] 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.
[56] 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.

**Mingwei Zhang** received the B.E. degree in automation from Zhengzhou University, Zhengzhou, China, in 2021. He is currently pursing the M.S. degree with the Unmanned System Research Institute and the school of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include remote sensing image acquisition and processing.

**QiangLi** received the Ph.D. degree in computer science and technology from Northwestern Polytechnical University, Xi'an, China in 2022. He is currently a postdoc with the School of Electronic Engineering, Xidian University, Xi'an. His research interests include remote sensing image processing and computer vision.



**Yanling Miao** received the B.E. degree in communication engineering and the M.S. degree in computer application technology from Henan Polytechnic University, Jiaozuo, China, in 2015 and 2019 respectively. She is currently pursuing the Ph.D. degree with the School of Computer Science and the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. Her research interests include hyperspectral image processing and computer vision.

**Yuan Yuan** (M'05-SM'09) is currently a Full Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. She has authored or co-authored over 150 papers, including about 100 in reputable journals, such as the IEEE TRANSACTIONS AND PATTERN RECOGNITION, as well as the conference papers in CVPR, BMVC, ICIP, and ICASSP. Her current research interests include visual information processing and image/video content analysis.

**Qi Wang** (M'15-SM'15) 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, machine learning, pattern recognition and remote sensing.