DGMA^2-Net: A Difference-Guided Multiscale Aggregation Attention Network for Remote Sensing Change Detection

Zilu Ying, Zijun Tan, Yikui Zhai, *Senior Member, IEEE*, Xudong Jia, Wenba Li, *Student Member, IEEE*, Junying Zeng, *Member, IEEE*, Angelo Genovese, *Senior Member, IEEE*, Vincenzo Piuri, *Fellow, IEEE*, and Fabio Scotti, *Senior Member, IEEE*

***Abstract*— Remote sensing change detection (RSCD) focuses on identifying regions that have undergone changes between two remote sensing images captured at different times. Recently, convolutional neural networks (CNNs) have shown promising results in the challenging task of RSCD. However, these methods do not efficiently fuse bitemporal features and extract useful information that is beneficial to subsequent RSCD tasks. In addition, they did not consider multilevel feature interactions in feature aggregation and ignore relationships between difference features and bitemporal features, which, thus, affects the RSCD results. To address the above problems, a difference-guided multiscale aggregation attention network, DGMA^2-Net, is developed. Bitemporal features at different levels are extracted through a Siamese convolutional network and a multiscale difference fusion module (MDFM) is then created to fuse bitemporal features and extract, in a multiscale manner, difference features containing rich contextual information. After the MDFM treatment, two difference aggregation modules (DAMs) are used to aggregate difference features at different levels for multilevel feature interactions. The features through DAMs are sent to the difference-enhanced attention modules (DEAMs) to strengthen the connections between bitemporal features and difference features and further refine change features. Finally, refined change features are superimposed from deep to shallow and a change map is produced. In validating the effectiveness of DGMA^2-Net, a series of experiments are conducted on three public RSCD benchmark datasets [LEVIR building change detection dataset (LEVIR-CD), Wuhan University building change detection dataset (BCDD), and Sun Yat-Sen University dataset (SYSU-CD)]. The experimental results demonstrate that DGMA^2-Net surpasses the current eight state-of-the-art methods in RSCD. Our code is released at https://github.com/yikuizhai/DGMA2-Net.**

***Index Terms*— Difference aggregation module (DAM), difference-enhanced attention module (DEAM), multiscale difference fusion module (MDFM), remote sensing change detection (RSCD).**

Manuscript received 4 February 2024; revised 7 April 2024; accepted 14 April 2024. Date of publication 17 April 2024; date of current version 25 April 2024. This work was supported in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2021A1515011576; in part by Guangdong Science and Technology Planning Project under Grant 2021A0505030080 and Grant 2021A0505060011; in part by Guangdong Higher Education Innovation and Strengthening School Project under Grant 2020ZDZX3031, Grant 2022ZDZX1032, and Grant 2023ZDZX1029; in part by Wuyi University Hong Kong and Macao Joint Research and Development Fund under Grant 2022WGALH19; and in part by Guangdong Jiangmen Science and Technology Research Project under Grant 2220002000246 and Grant 2023760300070008390. *(Zilu Ying, Zijun Tan, Yikui Zhai, and Xudong Jia contributed equally to this work.) (Corresponding authors: Yikui Zhai; Xudong Jia.)*

Zilu Ying, Zijun Tan, Yikui Zhai, Wenba Li, and Junying Zeng are with the Department of Intelligent Manufacturing, Wuyi University, Jiangmen 529020, China (e-mail: ziluy@163.com; 13555650396@163.com; yikuizhai@163.com; wenbalee@163.com; zengjunying@126.com).

Xudong Jia is with the College of Engineering and Computer Science, California State University at Northridge, Los Angeles, CA 18111 USA (e-mail: Xudong.Jia@csun.edu).

Angelo Genovese, Vincenzo Piuri, and Fabio Scotti are with the Department of Computer Science and the Dipartimento di Informatica, Università degli Studi di Milano, 20133 Milan, Italy (e-mail: angelo.genovese@unimi.it; vincenzo.piuri@unimi.it; fabio.scotti@unimi.it).

Digital Object Identifier 10.1109/TGRS.2024.3390206

I. Introduction

REMOTE sensing change detection (RSCD) is an important task to identify any changes occurring on the Earth's surface by analyzing bitemporal or multitemporal image acquired at the same geographic location. This task is playing an increasingly important role in disaster assessment [1], [2], environmental monitoring [3], land management [4], and urban transformation analysis [5]. Some samples of the RSCD task are shown in Fig. 1. Over the past few decades, researchers have done many studies on RSCD.

There are two main categories of traditional RSCD methods, pixel-based and object-based methods. The pixel-based methods calculate, transform, and compare each pixel of synchronized bitemporal or multitemporal images, and generate a change map through an appropriate threshold or a clustering algorithm. Commonly used methods include change vector analysis (CVA) [6], principal component analysis (PCA) [7], and slow feature analysis (SFA) [8]. However, these pixel-based methods only focus on a single pixel, ignoring embedded relationships between contexts in bitemporal or multitemporal images, resulting in many noises in change maps. To address this issue, scholars began to switch the research direction from pixels to objects. The object-based methods, such as Kullback-Leibler (KL) divergence [9] and image correlation analysis [10], focus on specific and low-level semantic objects and detect changes in bitemporal or multitemporal images. However, the low-level features in these methods are artificially created. They are inadequate in handling the complexity of real-world environments.

In recent years, deep learning has demonstrated remarkable performance in computer vision and has attracted great

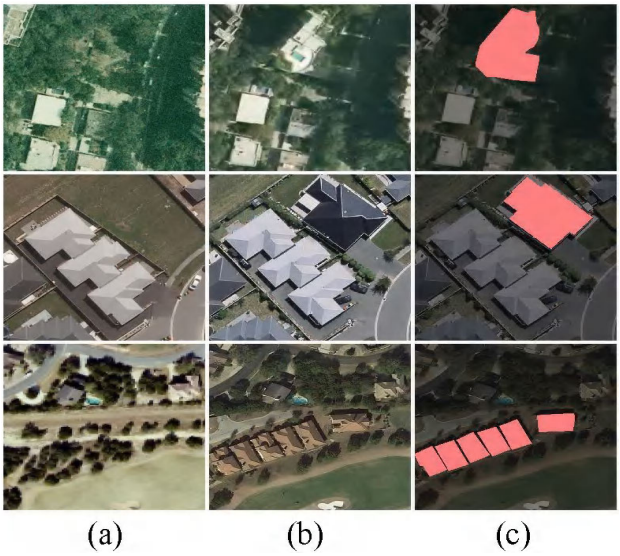

Fig. 1. RSCD examples. (a) T1 images. (b) T2 images. (c) Changes marked in red.

attention in various fields [11]. Convolutional neural networks (CNNs), thanks to their strong feature extraction abilities, have developed extensive applications in computer vision [12]. Example applications include but are not limited to image classification [13], [14], object detection [15], [16], and semantic segmentation [17], [18]. In the field of RSCD, CNNs have also emerged as a prominent research method for change detection and surpassed traditional RSCD methods. Weight-sharing Siamese CNNs are widely employed in deep learning-based RSCD methods [20] to extract features from bitemporal images. These methods are further categorized into metric-based methods [21], [22], [23], [24], [25], [26] and classification-based methods [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39]. Metric-based methods determine changes by computing parametric distances of bitemporal images in an embedding space where "pull in" distances are for changed pixels and "pull away" distances are for nonchanged pixels. Classification-based methods classify each pixel in the embedding space to get a set of changed regions.

Despite the favorable outcomes attained thus far in RSCD by current deep learning methods, significant challenges in feature fusion and aggregation persist in these methods. Feature fusion at each stage of the methods is limited to local feature interactions. In addition, most deep learning-based RSCD methods follow simple fusion operations, such as feature concatenation and feature subtraction [35], [40], [41]. Moreover, few studies have been concentrated on using multiscale feature leaning mechanisms in bitemporal feature fusion, although multiscale feature learning mechanisms have been proven to be effective in deep learning [54]. As a result, fused features often contain many noises, which adversely affect RSCD results.

Feature aggregation is also of great significance in RSCD. Features at different levels contain different pieces of semantic information. Shallow features contain low-level visual information, such as image texture, object edges, and other similar characteristics. With the network layers becoming deeper, these low-level features undergo a transition toward a higher level of abstraction, and deep features thus contain richer semantic information, such as image content and object location. U-Net models [42], [43], [44], [45], [46] are widely used in RSCD for feature aggregation. Within a U-Net network, deep features are superimposed with skip connections over shallow features to produce a change map. Furthermore, many RSCD works embed attention mechanisms into these methods to obtain more discriminative features [27], [28], [29], [30], [31]. These methods accomplish feature aggregation through feature concatenation, skip connections, and stepwise upsampling operations. However, these methods do not consider the multilevel feature interaction which plays a key role in RSCD [58]. Moreover, these methods ignore the relationships between difference features and bitemporal features, thus affecting the RSCD results.

To address the above-mentioned issues, a novel difference-guided multiscale aggregation attention network, DGMA^2-Net, is proposed. This network uses a multiscale feature learning mechanism to fuse bitemporal features and generate difference features. In detail, bitemporal features at different levels are extracted through a Siamese convolutional network and multiscale difference fusion modules (MDFMs) are then created to fuse bitemporal features. After the MDFM treatment, two difference aggregation modules (DAMs) are used to aggregate information at different depths and perform multilevel feature interactions. The features after the DAMs and the bitemporal features are sent into the difference-enhanced attention modules (DEAMs) to further enhance the change areas and refine difference features. A change map is produced by superimposing the refined difference features. The primary contributions of this work are outlined as follows.

1) An MDFM is proposed to fuse bitemporal features in multiscale manner. To the best of the authors' knowledge, MDFM is the first bitemporal feature module to take advantage of a multiscale feature learning mechanism to generate difference features and thus reduce noises in difference features.
2) A DAM is introduced to effectively aggregate difference features. Different from previous works that only focused on interactions within features, DAM explores the interactions in multilevel features for the aggregation of semantic information and makes full use of contexts at different levels.
3) A DEAM is proposed to further enhance change areas and refine difference features by revamping the self-attention mechanism. It is our knowledge that DEAM is the first one developed to investigate interactions between difference features and bitemporal features based on self-attention mechanism.
4) The integration of the MDFM, DAM, and DEAM in the DGMA^2-Net creates a new approach for RSCD. Quantitative and qualitative experimental results on three public datasets demonstrate that DGMA^2-Net achieves superior performance and exceeds state-of-the-art RSCD methods.

II. Related Works

A. Traditional RSCD Method

There have been many traditional RSCD methods. Among them, algebraic and clustering methods are representative methods. Algebraic methods include image difference, image ratio, and image regression [47], [48]. Combined with CVA [6] and SFA [8], algebraic methods are used to obtain different characteristics of bitemporal images and delineate change areas. Since these methods largely depend on the selection of a threshold, some transformation techniques, such as PCA [7] and tasseled cap transformation [49], are introduced to help determine the threshold and conduct RSCD. These threshold-based methods are subjective. In recognizing this limitation, clustering-based methods are brought in for RSCD [50], [51], [52]. However, these methods cannot understand the semantic meaning of changed regions, which leads to false positive errors. In addressing the above issues, object-based methods including KL divergence [9] and image correlation analysis [10] are developed to explore sematic interactions of changed objects.

It is noted that these methods can be easily implemented; however, they overlook the semantic and spatial information embedded in bitemporal images and are highly sensitive to noises. In addition, these methods require manual annotation of objects or features and rely on the arbitrary setting of a threshold. These methods are not suitable for RSCD in complex scenes.

B. Deep Learning-Based RSCD Method

Recently, the advancement in deep learning has led to a great success in developing many methods which use CNNs for RSCD. The deep learning-based methods can be categorized into metric-based methods [21], [22], [23], [24], [25], [26] and classification-based methods [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39].

1) Metric-Based Methods: These methods learn features and determine change regions or areas by using parameterized embedding distances of pixels in bitemporal images. These methods use contrastive loss [23], [25] or triplet loss [24] to optimize parameters (Params) of their models through training. Chen and Shi [22] introduced a spatiotemporal attention mechanism to catch features in bitemporal images and developed a new loss function to optimize the entire metric-based learning model. Liu et al. [25] used a denoising self-encoder and a probability model to eliminate noises, calculate distances between features in heterogeneous images, screen features by a preset threshold, and generate change maps.

2) Classification-Based Method: These methods fuse features in bitemporal images and group fused features into a feature space for classification. Li et al. [19] developed a temporal feature interaction module and used the module for refining and improving the detection quality of different remote sensing images. As a result, the efficiency of RSCD is achieved. Lei et al. [34] constructed a reverse attention mechanism and a multipart feature fusion strategy for RSCD. Fang et al. [35] constructed a densely connected Siamese network to reduce the information loss of a neural network through the transmission between an encoder and a decoder. Lei et al. [38] introduced a new RSCD network with a difference enhancement module and constructed an attention mechanism through spatial multiscale smooth pooling techniques, which effectively enhances edge integrity and internal compactness of changed regions.

It is noted that the effectiveness of the deep learning-based RSCD methods is strongly dependent on the quality of their acquired feature representations. Significant endeavors, therefore, should be dedicated to acquiring discriminative feature representations that optimize the performance of RSCD.

C. Bitemporal Feature Fusion

Bitemporal feature fusion is an essential process in RSCD models. The purpose of bitemporal feature fusion is to fuse or compare bitemporal features and capture their differences so that the model can better understand the change trends and extract difference features. Feature concatenation, feature subtraction, feature addition, and a series of attention mechanisms are used in feature fusion for RSCD. Fang et al. [35] fused bitemporal features through feature concatenation. Lei et al. [38] proposed a difference enhancement module in feature fusion operations, while Song et al. [39] introduced a context change enhancement module to leverage the quality of change detection from bitemporal images. Raza et al. [42] had a fusion operation through a parallel attention mechanism. All these studies constructed a channel attention mechanism through a set of global average pooling, maximum pooling, and feature subtraction operations for better feature fusion. It is important to point out that these RSCD methods focus on differences in bitemporal features. Using difference features to guide bitemporal feature fusion can be achieved better results. As an RSCD example with difference features, Li et al. [19] constructed a temporal feature interaction module which uses difference features and a series of arithmetic operations for bitemporal feature fusion. It is noted that although these methods have achieved satisfactory results, most of them use simple fusion operations or rely on existing attention mechanisms to achieve bitemporal feature fusion.

D. Feature Aggregation

Existing RSCD methods have successfully adopted U-Net, an encoder–decoder U-shaped architecture with skip connections, for feature aggregation. An example method developed by Daudt et al. [20] used three distinct U-Net-based modules for feature aggregation. However, these simple feature-based methods are difficult in performing multilevel feature interactions, thus affecting the RSCD results. Considering this issue, recent works have begun to introduce attention mechanisms for feature aggregation. Zhao et al. [36] built a triple-stream network which uses a self-attention mechanism along with skip connections and upsampling operations for feature aggregation. Zhang et al. [37] built a decoder with self-attention and channel-attention mechanisms and aggregated difference features and bitemporal features through concatenation and upsampling operations. Eftekhari et al. [27] also used a

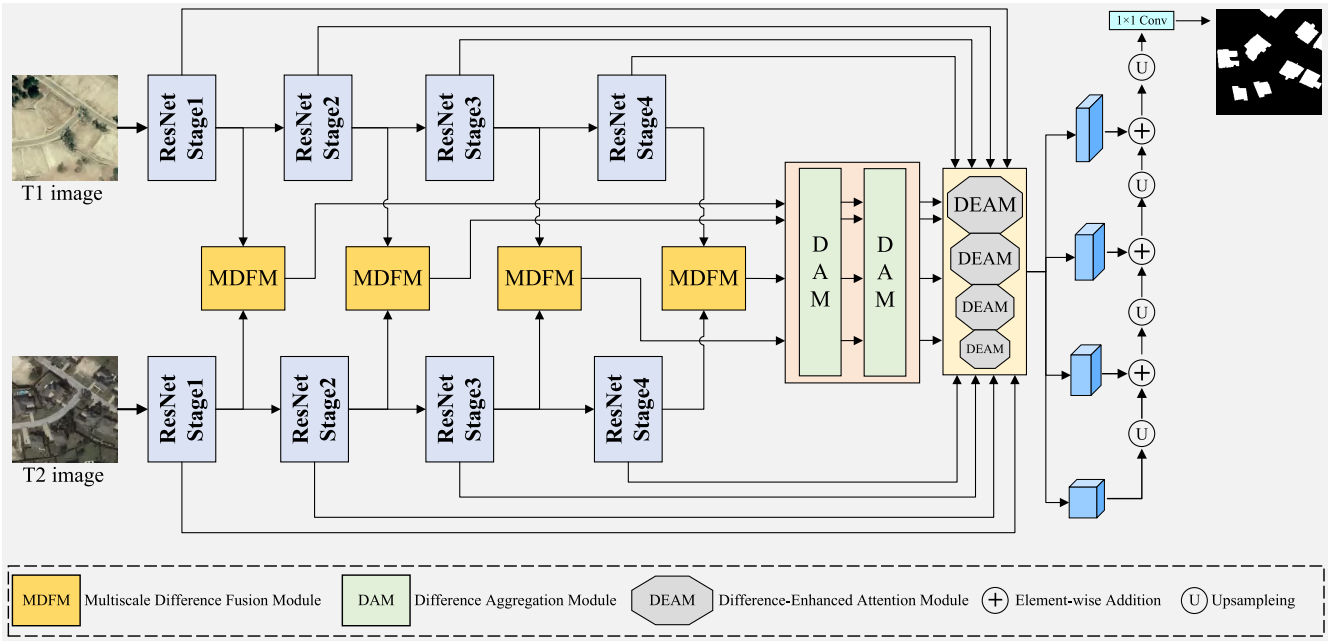

Fig. 2. Architecture of the DGMA^2-Net. Bitemporal images pass through a Siamese convolutional network to generate multilevel bitemporal features. These features are entered into the MDFMs to create difference features. Difference features are generated and refined through two DAMs and four DEAMs, and the resulting features are upsampled to create a change map.

self-attention mechanism to construct channel and spatial attentions and performed feature aggregation on bitemporal features through concatenation and upsampling operations. Lei et al. [30] constructed an attention mechanism through a Gaussian kernel, progressively upsampled difference features, and used skip connections to complete feature aggregation. It is noted that these methods with skip connections lack capabilities of interacting multilevel features. As a result, they do not efficiently aggregate difference features with complementary information of contexts in bitemporal features. It is also noted that self-attention mechanisms have been applied on bitemporal features or difference features in RSCD. However, these self-attention mechanisms do not consider the coupling interactions of bitemporal features and difference features, thus affecting the RSCD results.

III. Method

A. DGMA^2-Net Architecture

This section describes the DGMA^2-Net architecture. As shown in Fig. 2, DGMA^2-Net comprises an encoder and a decoder for RSCD. The encoder first uses the pretrained Siamese weight-sharing ResNet-18 [53] to extract multilevel bitemporal features from a pair of registered and synchronized bitemporal images (T1 and T2), and then utilizes the MDFMs to fuse the extracted features. After the MDFM operations, difference features along with their context information are further fed to the DAMs for feature aggregation and interaction. The refined features are then sent to the DEAMs to further enhance the changes. A change map is produced by superimposing the feature from deep to shallow.

B. Multiscale Difference Fusion Module (MDFM)

In DGMA^2-Net, the MDFM fuses features obtained from bitemporal images and generates difference features with

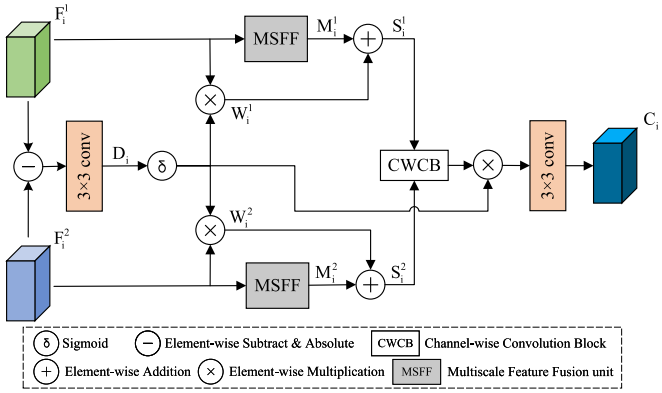

Fig. 3. MDFM structure.

valuable context information. In Fig. 3, bitemporal features coming from two paired images are denoted by F_i^1 and F_i^2, where $i \in \{1, 2, 3, 4\}$, represents the ResNet-18 stage index. F_i^1 and F_2^1, extracted by the weight-sharing ResNet-18, go through a series of pixel-by-pixel subtraction, absolute value operation, and 3 × 3 convolution operation. The initial difference feature D_i is then generated. This process can be expressed as follows:

$$D_i = \mathrm{Conv}_{3\times3}\left(\left|F_i^1 - F_i^2\right|\right) \tag{1}$$

where $|\cdot|$ represents an absolute operation and $\boldsymbol{Conv_{3\times3}}(\cdot)$ represents a 3 × 3 convolutional layer, a batch normalization layer, and a rectified linear unit (ReLU) activation function.

It is important to point out that after the initial difference feature D_i is generated, the multiscale feature learning mechanism created by Szegedy et al. [54] is enhanced in this study for fusing bitemporal features. Using the convolutions of different kernel sizes, the MDFM constructs a multiscale fusion process. As shown in Fig. 4, the MDFM employs a pair of multiscale feature fusion (MSFF) units to enhance the

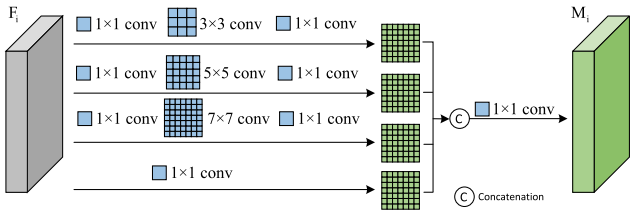

Fig. 4. MSFF unit structure.

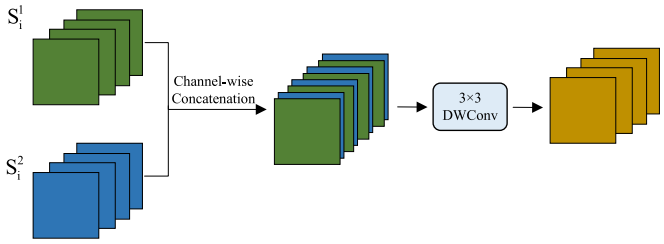

Fig. 5. CWCB structure.

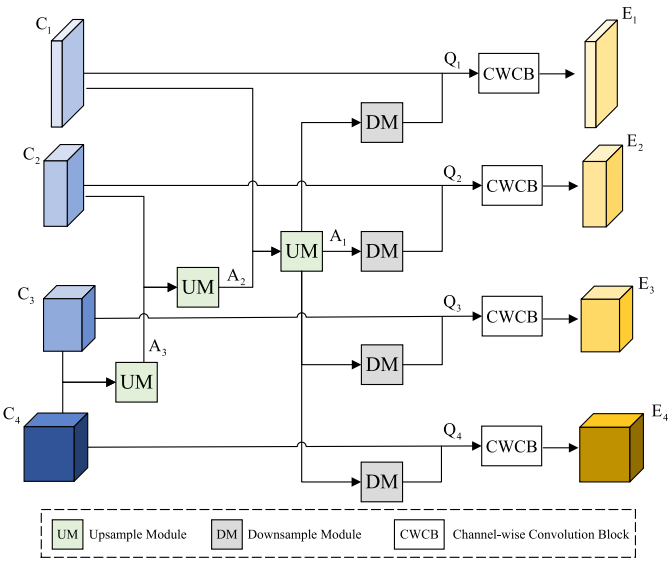

Fig. 6. DAM structure.

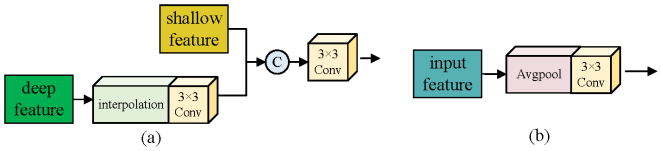

Fig. 7. (a) UM and (b) DM structure.

fusing operations. The MSFF unit includes four branches of convolutional operations with different kernel sizes, three of which contain $[1 \times 1, 3 \times 3, 1 \times 1]$, $[1 \times 1, 5 \times 5, 1 \times 1]$, and $[1 \times 1, 7 \times 7, 1 \times 1]$ convolutional fusing operations. The fourth branch only contains a 1×1 convolution operation. The MSFF process can be described as follows:

$$\begin{aligned} M_i = &\ \mathrm{Conv}_{1\times1}(\mathrm{Cat}(\mathrm{Conv}_{1\times1}(F_i))) \\ &\ \mathrm{Conv}_{1\times1}(\mathrm{Conv}_{3\times3}(\mathrm{Conv}_{1\times1}(F_i))) \\ &\ \mathrm{Conv}_{1\times1}(\mathrm{Conv}_{5\times5}(\mathrm{Conv}_{1\times1}(F_i))) \\ &\ \mathrm{Conv}_{1\times1}(\mathrm{Conv}_{7\times7}(\mathrm{Conv}_{1\times1}(F_i))) \end{aligned} \tag{2}$$

where $\boldsymbol{Cat}(\cdot)$ is a concatenation operation and M_i is the concatenated features of the four branches.

An elementwise channel weight W_i is introduced in DGMA^2-Net. Following the idea of residual, S_i, which simultaneously incorporate multiscale information, is generated by adding W_i and M_i. This process can be expressed as follows:

$$W_i = F_i \times \sigma(\mathrm{Conv}_{3\times3}(D_i)) \tag{3}$$

$$S_i = W_i + M_i. \tag{4}$$

The channelwise convolution block (CWCB) is employed in DGMA^2-Net to perform the bitemporal feature fusion operation. As shown in Fig. 5, S_i^1 and S_i^2 are concatenated channel-by-channel and the concatenated features are operated through a 3×3 depthwise convolution operation. The fused features through CWCB are multiplied by W_i. The final fused difference feature C_i is obtained after the 3×3 convolution operation is completed. C_i can be expressed as follows:

$$C_i = \mathrm{Conv}_{3\times3}\left(\mathrm{DWConv}_{3\times3}\left(\mathrm{Stack}\left(S_i^1, S_i^2\right)\right) \times W_i\right) \tag{5}$$

where $\boldsymbol{DWConv}(\cdot)$ represents a depthwise convolution operation and $\boldsymbol{Stack}(\cdot)$ represents the channelwise feature concatenation operation.

C. Difference Aggregation Module (DAM)

The DAM is developed in this study to take the advantage of feature differences and effectively aggregate features by considering multilevel interactions of difference features. As shown in Fig. 6, the upsample modules (UMs) are used to fuse deep features with shallow features; that is, deep features from their previous ResNet-18 stage are transformed to be the same size as shallow features (or the fused features from the previous ResNet-18 stage). After the transformation, a 3×3 convolution operation is applied on deep features [see Fig. 7(a)].

The deep features are further concatenated with the shallow features and a 3×3 convolution operation is conducted again. The aggregation process can be mathematically expressed as follows:

$$A_{i-1} = \mathrm{Conv}_{3\times3}(\mathrm{Cat}(\mathrm{shallow}, \mathrm{Conv}_{3\times3}(\mathrm{UP}(\mathrm{deep})))) \tag{6}$$

where shallow represents the shallow feature, deep represents the deep feature and $\boldsymbol{UP}(\cdot)$ represents the interpolation operation.

After all the features are aggregated, the final aggregated feature A_1 needs to be reverted to the same size as C_i, and it is sent into the downsample module (DM) [see Fig. 7(b)]. Specifically, A_1 is reverted by the average pooling at different scales, after a 3×3 convolution operation is executed. The refined feature Q_i is obtained through the following calculation:

$$Q_i = \begin{cases} \mathrm{Conv}_{3\times3}(A_1), & i = 1 \\ \mathrm{Conv}_{3\times3}(\mathrm{Avgpool}(\mathrm{Conv}_{3\times3}(A_1))), & \text{other} \end{cases} \tag{7}$$

where $\boldsymbol{Avgpool}(\cdot)$ represents an average pooling operation.

It is pointed out that skip connections are applied in each DAM to prevent the loss of semantic information. Moreover,

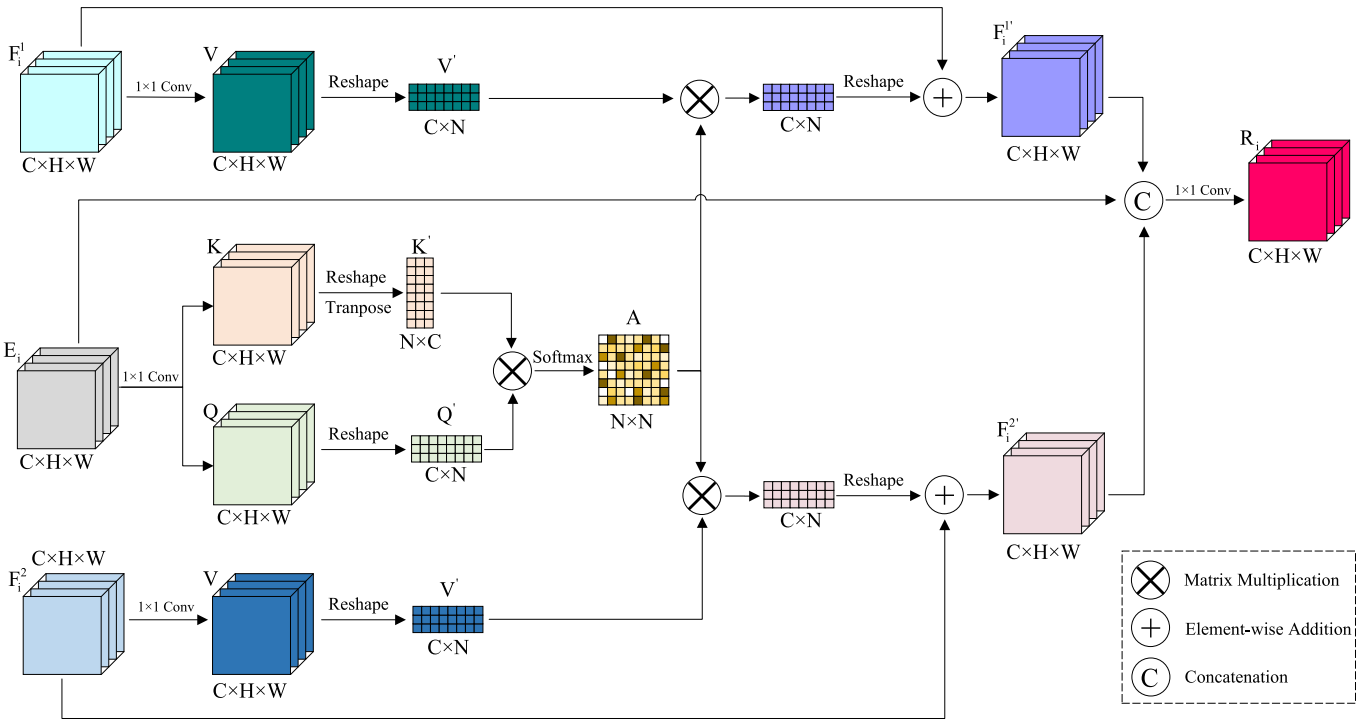

Fig. 8. DEAM structure.

CWCB is also used to utilize the multilevel difference features C_i and the refined features Q_i to obtain the refined difference features E_i. To achieve the best result, two DAMs are applied in DGMA2-Net and the ablation experiments described in Section IV confirm the necessity of two DAMs.

D. Difference-Enhanced Attention Module (DEAM)

In DGMA^2-Net, a set of DEAMs are used for linking difference features to bitemporal features (see Fig. 2). Each DEAM expands the structure of a self-attention mechanism by building query (Q), key (K), and value (V). This module demonstrates an innovative approach to establishing the relationships between difference features and bitemporal features. In detail, four DEAMs are embedded in DGMA^2-Net to receive difference features E_i from the two DAMs and mine the relationships between difference features and bitemporal features (F_i^1 and F_i^2). For each DEAM, a set of attention operations are conducted (see Fig. 8).

A 1×1 convolution is conducted on the refined difference feature $E_i \in \mathbb{R}^{C\times H\times W}$ and E_i is split into two feature vectors, key ($K \in \mathbb{R}^{(C/8)\times H\times W}$) and query ($Q \in \mathbb{R}^{(C/8)\times H\times W}$). After K and Q are transformed, matrix multiplication is performed to obtain attention matrix $A \in \mathbb{R}^{N\times N}$.

A 1×1 convolution is also applied to $F_i^1 \in \mathbb{R}^{C\times H\times W}$ and $F_i^2 \in \mathbb{R}^{C\times H\times W}$, respectively. The result of the convolution is noted as the value vector $V \in \mathbb{R}^{C\times H\times W}$ for the two bitemporal features (F_i^1 and F_i^2), respectively. The value vector V is further reshaped to be V' for bitemporal features.

To further integrate difference features and bitemporal features, the DEAM multiplies the attention matrix A by the V' vector. The result of the multiplication is reshaped and summed with F_i^1 and F_i^2, respectively, to obtain the difference-enhanced features $F_i^{1'}$ and $F_i^{2'}$. This process can be expressed as follows:

$$F_i^{1'} = V' \otimes A + F_i^1 \tag{8}$$

$$F_i^{2'} = V' \otimes A + F_i^2 \tag{9}$$

where $\otimes$ represents the matrix multiplication.

$F_i^{1'}$, $F_i^{2'}$, and E_i are further concatenated and processed by a 1×1 convolution kernel to get a difference-enhanced feature R_i, which can be formulated as follows:

$$R_i = \mathrm{Conv}_{1\times 1}\left(\mathrm{Cat}\left(F_i^{1'}, F_i^{2'}, E_i\right)\right). \tag{10}$$

The difference-enhanced features are superimposed from shallow to deep, step by step, to create the final feature, and a change map is obtained by applying a 1×1 convolution layer on the final feature.

E. Hybrid Loss Function

In recognizing that unchanged areas prevail in bitemporal images [35], the DGMA^2-Net employs a hybrid loss function to optimize the training process. The hybrid loss function combines the cross-entropy loss function and the dice loss function. The cross-entropy loss function is formulated as follows:

$$L_{\mathrm{BCE}} = \frac{1}{H\times W}\sum_{k=1}^{H\times W} y_k \log\left(\hat{y}_k\right) + (1-y_k)\log\left(1-\hat{y}_k\right) \tag{11}$$

where H and W represent the height and width of bitemporal images, respectively, $\hat{y}_k$ represents the prediction result of the kth pixel in bitemporal images, and y_k represents the

ground-truth associated with the kth pixel. The dice loss function can be formulated as follows:

$$L_{\text{DICE}} = 1 - \frac{2\sum_{k=1}^{H\times W} y_k \widehat{y}_k}{\sum_{k=1}^{H\times W}\left(y_k + \widehat{y}_k\right)}. \tag{12}$$

The hybrid loss function can be formulated as follows:

$$L_{\text{TOTAL}} = L_{\text{BCE}} + L_{\text{DICE}}. \tag{13}$$

IV. Experiment

A. Experimental Setup

In this study, a set of experiments were conducted to verify the effectiveness of DGMA^2-Net.

1) Dataset: Three public RSCD datasets, namely, LEVIR building change detection dataset (LEVIR-CD) [22], Wuhan University building change detection dataset (WHU-BCDD), (BCDD) [41], and Sun Yat-Sen University dataset (SYSU-CD) [21] were used in the experiments.

The LEVIR-CD dataset consists of 637 pairs of images obtained from the Google Earth application programming interface (API). Each image has a size of 1024 × 1024 pixels and a resolution of 0.5 m. This dataset primarily focuses on changes in physical buildings. The dataset is randomly partitioned into training, testing, and validation sets by a ratio of 7:1:2. All images were cropped to be 256 × 256 pixels. A total of 7120 pairs of images were available for training, 1024 pairs for validation, and 2048 pairs for testing.

The BCDD dataset consists of a pair of aerial images with 32 507 × 15 354 pixels. The resolution of the images is 0.075 m. Following the guidance from bitemporal image transformer (BIT) [55], we cropped the aerial images into 7620 pairs of 256 × 256 blocks without overlapping and randomly made the training set with 6096 pairs, the validation set with 692 pairs, and the test set with 692 pairs.

The SYSU-CD dataset consists of 20 000 pairs of images with a size of 256 × 256 pixels and a resolution of 0.5 m. It encompasses diverse and intricate change scenes such as expanded roads, new urban buildings, vegetation changes, and enlarged suburban areas. The dataset is randomly partitioned in this study into 12 000 pairs of images for training, 4000 pairs for validation, and 4000 pairs for testing.

2) Evaluation Metrics: Five widely used metrics, namely overall accuracy (OA), precision (Pre), recall (Rec), $F1$ score ($F1$), and intersection over union (IoU), were adopted in this study for evaluating the performance of DGMA2-Net. These metrics are defined as follows:

$$\text{OA} = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \tag{14}$$

$$\text{precision} = \frac{T_p}{T_p + F_p} \tag{15}$$

$$\text{recall} = \frac{T_p}{T_p + T_n} \tag{16}$$

$$F1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{17}$$

$$\text{IoU} = \frac{T_p}{T_p + F_p + F_n} \tag{18}$$

where T_p represents the sum of accurately classified changed pixels, F_p represents the sum of truly unchanged pixels misclassified as changed, T_n represents the sum of accurately classified unchanged pixels, and F_n represents the sum of unchanged pixels falsely detected as changed. OA denotes the proportion of accurately classified pixels to the total pixels. The precision measures the false positive rate of the DGMA^2-Net, while recall reflects the false negative rate. $F1$, which considers both precision and recall, provides a comprehensive measure of the model performance. IoU quantifies the degree of overlapping changes on the detection map and the ground-truth map, making it a crucial metric for evaluating the accuracy of the DGMA^2-Net.

3) Implementation Details: PyTorch framework was adopted in this study for all experiments and a single GeForce RTX 2080ti was used for accelerating the model training process. Random flipping and cropping techniques were employed to augment images from the three datasets. The DGMA^2-Net was trained using the Adam optimizer with a momentum of 0.9 and a weight decay of 0.0001. In addition, β_1 and β_2 of Adam optimizer were set to be 0.9 and 0.99, respectively. During training, the batch size was set to be 16, and the learning rate was set to be 0.0001. In addition, a poly learning rate decay strategy [19], that is, $(1 - \text{cur\_iteration}/\text{max\_iteration})^{\text{power}} \times \text{lr}$, was followed, where power and max_iteration were set to be 0.9 and 80 000, respectively.

B. Comparative Models

Three traditional methods, CVA [6], SFA [8], PCA-Kmeans [50], and eight state-of-the-art deep learning-based methods, FC-Conc [20], FC-Diff [20], Siamese network and NestedUNet (SNUNet) [35], BIT [55], geospatial-awareness network (GeSANet) [56], dual-branch multilevel intertemporal network (DMINet) [29], intra-scale cross-interaction and inter-scale feature fusion network (ICIF-Net) [57], and Ultralightweight Spatial–Spectral Feature Cooperation Network (USSFC-Net) [30], were used to be the baseline models for comparing with DGMA^2-Net on the three datasets (see Section IV-A). These eight methods are briefly described with the following highlights.

1) *CVA:* It calculates the change vector of each pixel in bitemporal images and obtains the change areas through threshold filtering.
2) *SFA:* It learns transformations to extract slow features from bitemporal images. The learned slow features can be used to represent changes in each pixel position.
3) *PCA-Kmeans:* It applies PCA and K-means clustering on difference images for change detection. Each pixel pair, corresponding to identical positions in distinct images, is transformed into a feature vector.
4) *FC-Conc:* It employs a Siamese neural network and a feature-level fusion approach to extract multilevel concatenated features.
5) *FC-Diff:* Like FC-Conc, it employs a feature fusion method to leverage the absolute difference between feature pairs in the fusion operations of bitemporal images.
6) *SNUNet:* It takes advantage of the Siamese network and the Nested-UNet architecture and adopts a multilevel feature connection method. It applies a channel attention mechanism to each layer of decoded features and

TABLE I

QUANTITATIVE ANALYSIS OF RSCD MODELS (PRECISION, RECALL, $F1$, IOU, AND OA) ON THREE DATASETS. THE BEST VALUES ARE IN BOLD. ALL RESULTS ARE DESCRIBED IN PERCENTAGES (%)

Dataset		LEVIR-CD					BCCD					SYSU-CD				
		Pre	Rec	F1	IoU	OA	Pre	Rec	F1	IoU	OA	Pre	Rec	F1	IoU	OA
Traditional Methods	CVA [6]	5.040	61.36	9.320	4.890	39.19	4.130	62.96	7.750	4.030	38.85	24.28	62.13	34.92	21.15	45.39
	SFA [8]	6.110	36.32	10.45	5.510	68.28	5.910	44.18	10.43	5.500	69.03	31.83	41.61	36.06	22.00	65.21
	PCA-Kmeans [50]	5.160	37.07	9.050	4.740	62.06	6.410	48.53	11.32	6.000	69.08	43.18	46.80	44.92	28.96	72.93
Deep Learning Methods	FC-Conc [20]	82.03	63.44	71.55	55.70	97.43	71.81	64.23	67.81	51.29	97.51	76.99	61.88	68.61	52.22	86.65
	FC-Diff [20]	88.22	68.31	77.00	62.60	97.92	66.25	62.99	64.58	47.69	97.18	78.55	45.72	57.80	40.65	84.25
	SNUNet [35]	91.06	87.74	89.37	80.78	98.94	87.75	73.52	80.01	66.68	98.50	81.60	81.61	81.60	68.93	91.32
	BIT [55]	90.69	88.45	89.56	81.09	98.95	94.98	92.05	93.49	87.78	99.48	81.69	77.82	79.71	66.26	90.66
	GeSANet [56]	91.31	88.94	90.11	82.00	99.01	86.59	71.97	78.61	64.75	98.40	75.11	82.57	78.66	64.83	89.44
	DMINet [29]	87.74	**92.75**	90.18	82.11	98.97	90.18	78.55	83.97	72.36	98.78	80.05	82.89	81.44	68.70	91.09
	ICIF-Net [57]	86.08	92.23	89.05	80.26	98.84	90.90	82.41	86.45	76.13	98.95	76.69	80.69	78.64	64.80	89.66
	USFFC-Net [30]	89.37	91.53	90.44	82.54	99.04	87.58	**95.75**	91.48	84.30	99.33	80.41	78.72	79.56	66.05	90.25
	DGMA^2-Net	**92.13**	90.31	**91.21**	**83.84**	**99.14**	**96.26**	92.77	**94.48**	**89.54**	**99.56**	**84.32**	**83.24**	**83.78**	**72.08**	**92.40**

employs a deep supervision mechanism to enhance the ability of discriminating features.

7) *BIT:* It employs a transformer encoder to capture contextual information within a condensed tokenized space–time framework by representing bitemporal images as multiple label tokens. Subsequently, the extracted tokens, enriched with global information, are reintroduced into the embedding space. In doing so, original features are enhanced through the transformer decoder.
8) *GeSANet:* It is a model based on deformable convolution and the CANDECAMP/PARAFAC decomposition theory. It uses deformable convolution to conduct the pixel-level adaptive matching of multitemporal images and uses CANDECAMP/PARAFAC decomposition theory to reconstruct tensors to filter pseudochange information.
9) *DMINet:* It unifies a self-attention mechanism and an across-attention mechanism in a single module to steer the global feature distribution. It uses subtraction and concatenation as well as multilevel difference aggregation for detecting changes.
10) *ICIF-Net:* It integrates CNN and transformer to facilitate the interactions of global and local features. In addition, an attention mechanism is constructed through convolutions to integrate information from features of different resolutions.
11) *USSFC-Net:* It is a lightweight model which implements a new multiscale feature extraction method and constructs an attention mechanism through the Gaussian kernel. It relies on spatial–spectral dependences of extracting richer features.

All training Params except the batch size were set to be the default Params of the above models. Due to the limitation of memory size, the batch size of SNUNet and USFFC-Net was adjusted to be 16 in the study.

C. Comparisons of DGMA^2-Net With Baseline RSCD Models

In this study, DGMA^2-Net was compared with the 11 baseline models. It is noted that $F1$ and IoU are the metrics in RSCD that best reflect model performance so that the quality of a model is ultimately judged through $F1$ and IoU. The quantitative analysis of the models on the three datasets is shown in Table I and the below are our findings.

1) Analysis of Traditional Methods: The results from Table I clearly indicate that the performance metrics of traditional methods are significantly lower than those of deep learning-based methods on the three datasets. For instance, on the LEVIR-CD dataset, the highest $F1$ achieved by any of the three traditional methods is only 10.45, whereas the lowest $F1$ achieved by the deep learning-based method is 71.55. Visual results of each method are also demonstrated in Figs. 9–11. It can be observed from columns (d)–(f) that the detection results of traditional methods exhibit considerable noise, far inferior to those achieved by deep learning methods. These findings indicate the challenges faced by traditional methods in complex environments and underscore the superior performance of deep learning in RSCD.

2) Analysis of Deep Learning-Based Methods:

a) Experiments on LEVIR-CD: The RSCD models' performances in predicting changes in bitemporal images from the LEVIR-CD dataset are shown in Fig. 9. In these models, it is obvious that DGMA^2-Net, compared with the baseline models, is effective in RSCD. As shown in Table I, although DGMA^2-Net is not the best model in terms of recall, it is the best from the view of other metrics [precision (0.9213), $F1$ (0.9121), IoU (0.8384), and OA (0.9914)]. In contrast, the DMINet model has the highest recall (0.9275) among the RSCD models, slightly higher than DGMA^2-Net (0.9031). However, the precision of the DMINet model is relatively low (0.8774), which indicates that the DMINet model has more false positives.

b) Experiments on BCDD: The RSCD models' performances in predicting changes in bitemporal images from the BCDD dataset are shown in Fig. 10. It is noted that DGM2A-Net has fewer missed detections and less false positive errors than the baseline models, when it deals with detections of physical buildings in a relatively large area. For physical buildings with small areas, DGMA^2-Net is successful in extracting the buildings completely and identifying the edges

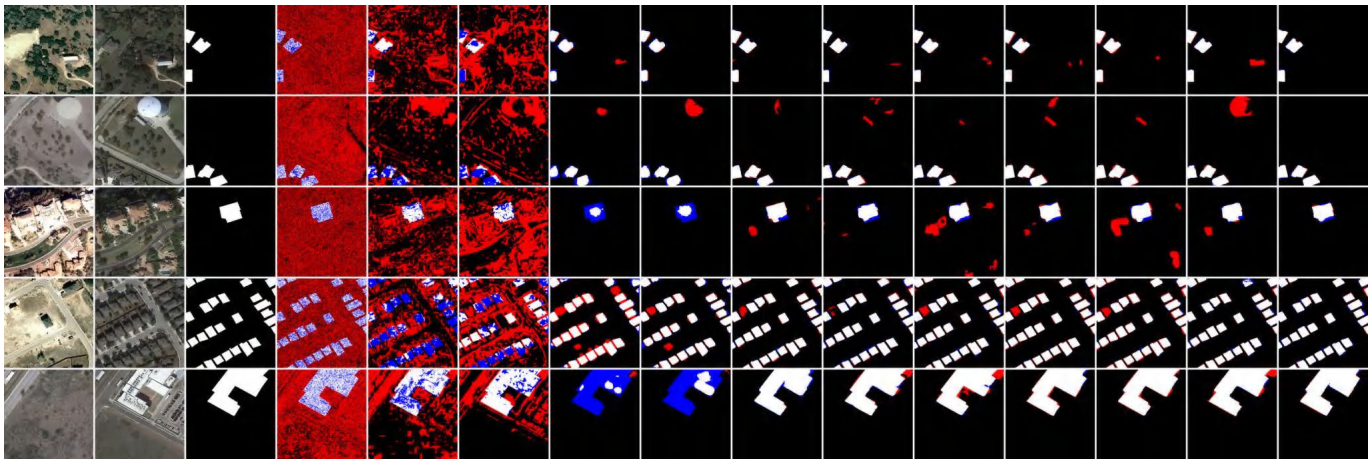

Fig. 9. Visual comparisons of DGMA^2-Net and the baseline RSCD models on the LEVIR-CD dataset. (a) T1 images, (b) T2 images, (c) ground truth, (d) CVA, (e) SFA, (f) PCA-Kmeans, (g) FC-Conc, (h) FC-Diff, (i) SNUNet, (j) BIT, (k) GeSANet, (l) DMINet, (m) ICIF-Net, (n) USFFC-Net, and (o) DGMA^2-Net.

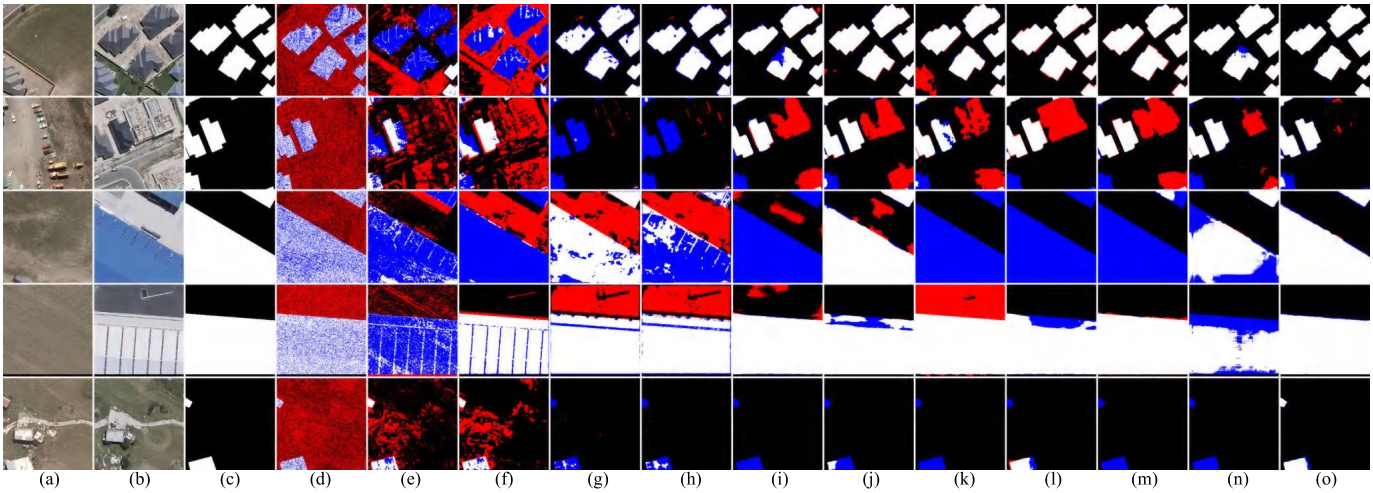

Fig. 10. Visual comparisons of DGMA^2-Net and the baseline models on the BCDD dataset. (a) T1 images, (b) T2 images, (c) ground truth, (d) CVA, (e) SFA, (f) PCA-Kmeans, (g) FC-Conc, (h) FC-Diff, (i) SNUNet, (j) BIT, (k) GeSANet, (l) DMINet, (m) ICIF-Net, (n) USFFC-Net, and (o) DGMA^2-Net.

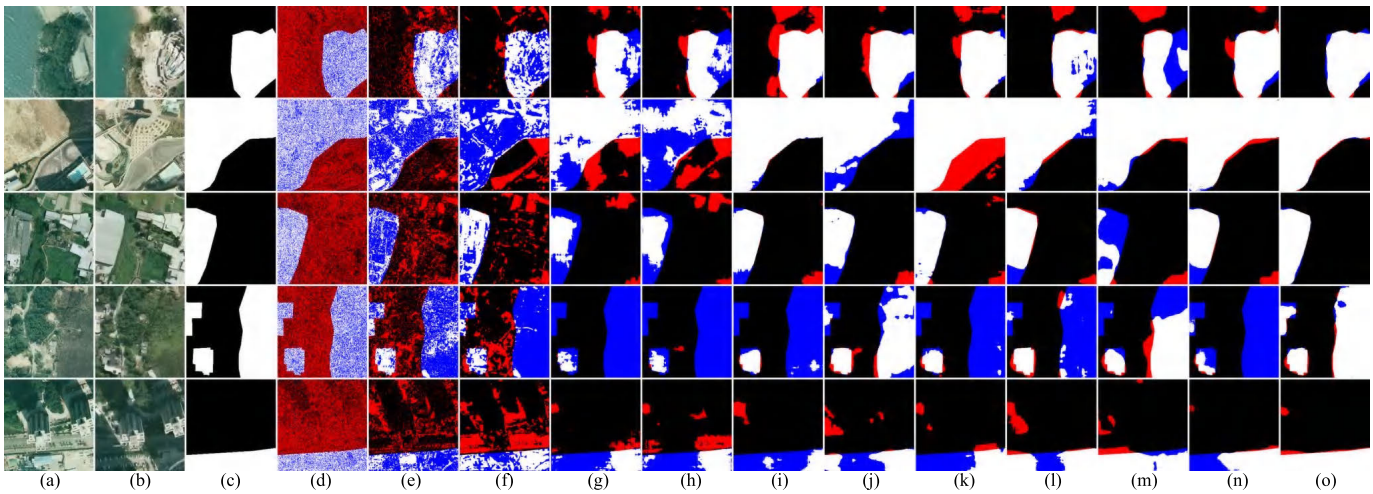

Fig. 11. Visual comparisons of DGMA^2-Net and the baseline models on SYSU-CD dataset. (a) T1 images, (b) T2 images, (c) ground truth, (d) CVA, (e) SFA, (f) PCA-Kmeans, (g) FC-Conc, (h) FC-Diff, (i) SNUNet, (j) BIT, (k) GeSANet, (l) DMINet, (m) ICIF-Net, (n) USFFC-Net, and (o) DGMA^2-Net.

of the buildings, while the baseline models have noticeable false and missed detections. Moreover, it can be seen from Table I that the DGMA^2-Net is superior to the baseline models in precision (0.9626), $F1$ (0.9448), IoU (0.8954), and OA (0.9956). The USFFC-Net model has the highest recall (0.9575), higher than DGMA^2-Net (0.9277). However, the precision of the DMINet model is low (0.8758), which implies many false positives.

TABLE II

QUANTITATIVE ANALYSIS OF MODELS ON PARAMS, FLOPS, TRAIN TIME FOR ONE EPOCH ON LEVIR-CD DATASET, AND FPS ON LEVIR-CD DATASET. THE BEST VALUES ARE IN BOLD

Method	Params(M)	Flops(G)	Epoch time(min)	FPS
FC-Conc [20]	1.55	5.32	**1.0**	52
FC-Diff [20]	**1.35**	**4.72**	**1.0**	56
SNUNet [35]	12.03	54.82	9.2	37
BIT [55]	3.01	8.48	1.5	40
GeSANet [56]	36.50	8.61	2.3	10
DMINet [29]	6.24	14.48	3.1	**61**
ICIF-Net [57]	25.30	23.84	7.0	16
USFFC-Net [30]	1.52	4.86	2.0	16
DGMA^2-Net	37.10	18.10	2.9	22

c) Experiments on SYSU-CD: It is noteworthy that all metrics of the models in the SYSU-CD dataset are lower than those in the LEVIR-CD and BCDD datasets (see Table I), because the change objects in the SYSU dataset are more complex, and the detection results are more susceptible to noise, which leads to more false detections. Among the models, DGMA^2-Net showcases its superior performance in all metrics. It has its precision, recall, $F1$, IoU, and OA to be 0.8432, 0.8324, 0.8378, 0.7208, and 0.9240, respectively. The RSCD models' performances in predicting changes in bitemporal images from the SYSU-CD dataset are shown in Fig. 11. It is observed that DGMA^2-Net is better in detecting change areas with fewer false and missed detections, while the baseline models have more false and missed detections.

In summary, the experiments on the three datasets confirmed the effectiveness of DGMA2-Net, which can be attributed to DGMA2-Net's capabilities of multiscale feature learning capabilities and multilevel feature interaction, as well as adequate operations of bitemporal features and difference features.

d) Analysis of model efficiency: Table II shows the number of Params, floating-point operations (Flops) of the deep learning-based models, as well as the time needed to run an epoch and frames per second (FPS) on LEVIR-CD dataset. It is noted that although the Params of DGMA^2-Net are higher than the baseline models, the Flops and training efficiency are not much different from the baseline models. Unfortunately, the FPS is relatively low among all the models, which indicates that exploring the potential for acceleration and investigating certain lightweight architectures is still necessary.

D. Ablation Experiment

The DGMA^2-Net is composed of three key components: MDFM, DAM, and DEAM. The ablation experiment in this study therefore is aimed at assessing the marginal impacts of these three components on DGMA^2-Net. Findings from the ablation study are as follows.

1) MDFM Ablation Experiment: In DGMA^2-Net, the MDFM is critical to fusing bitemporal features and embedding useful semantic information into difference features in a multiscale manner. This study used two standard fusing operations (concatenation and subtraction), DEM [31] and TFIM [21] as the ablation baseline modules to compare with the MDFM. The concatenation operation directly concatenates bitemporal features, while the subtraction operation gets the difference of bitemporal features and performs absolute value operations.

TABLE III

ABLATION EXPERIMENT ON MDFM. THE BEST VALUES ARE IN BOLD

Dataset	Feature Fusion	Pre	Rec	F1	IoU	OA
LEVIR-CD	Concatenation	91.12	89.12	90.11	82.00	99.00
	Subtraction	91.00	89.05	90.01	81.84	98.99
	DEM	90.63	**90.41**	90.52	82.68	99.04
	TFIM	91.34	89.84	90.59	82.79	99.05
	MDFM	**92.13**	90.31	**91.21**	**83.84**	**99.14**
BCDD	Concatenation	94.79	88.85	91.72	84.71	99.35
	Subtraction	92.44	90.66	91.54	84.40	99.32
	DEM	95.53	90.44	92.92	86.77	99.44
	TFIM	95.16	91.65	93.37	87.57	99.47
	MDFM	**96.26**	**92.77**	**94.48**	**89.54**	**99.56**
SYSU-CD	Concatenation	81.86	82.67	82.26	69.87	91.59
	Subtraction	85.93	79.92	82.82	70.68	92.18
	DEM	86.24	80.23	83.13	71.13	92.32
	TFIM	**86.46**	79.94	83.07	71.04	92.32
	MDFM	84.32	**83.24**	**83.78**	**72.08**	**92.40**

Table III presents the ablation experimental results. It is noted when MDFM is removed and replaced by concatenation or subtraction, the model performance decreases significantly. In comparison with concatenation and subtraction, DEM and TFIM have better performance, but the performance is still not as good as MDFM. Specifically, while the MDFM is the best in $F1$, IoU, and OA on the three datasets. In the LEVIR-CD dataset, the recall of the MDFM is 0.9031, only second to the DEM (0.9041). In the BCDD dataset, the MDFM achieves the best in all metrics. In the SYSU-CD dataset, the MDFM is the best in terms of recall, $F1$, IoU, and OA.

The visual results are shown in Fig. 12. When concatenation, subtraction, DEM, and TFIM are employed to conduct temporal feature fusion, the change results are degraded by the background noises, while our MDFM performs well in reducing mistaken and missed predictions. This confirms that the use of the concurrent convolution operations with different kernel sizes in the multiscale feature learning mechanism effectively alleviates the fusion of semantic information of bitemporal features into difference features and thus reduces noises.

2) DAM Ablation Experiment: The DAM is developed to aggregate features and perform multilevel feature interactions after the MDFM operations. Table IV lists the ablation experimental results. It is interesting to note that using one DAM is not the optimal configuration since the ability of one DAM in guiding the learning of features is limited. The optimal configuration is with two DAMs. When zero to four DAMs are considered in DGMA^2-Net, the resulting metrics ($F1$, IoU, and OA) indicate that DGMA^2-Net with two DAMs is the best choice. After two DAMs are embedded in DGMA^2-Net, the performance (in terms of $F1$, IoU, and OA) decreases

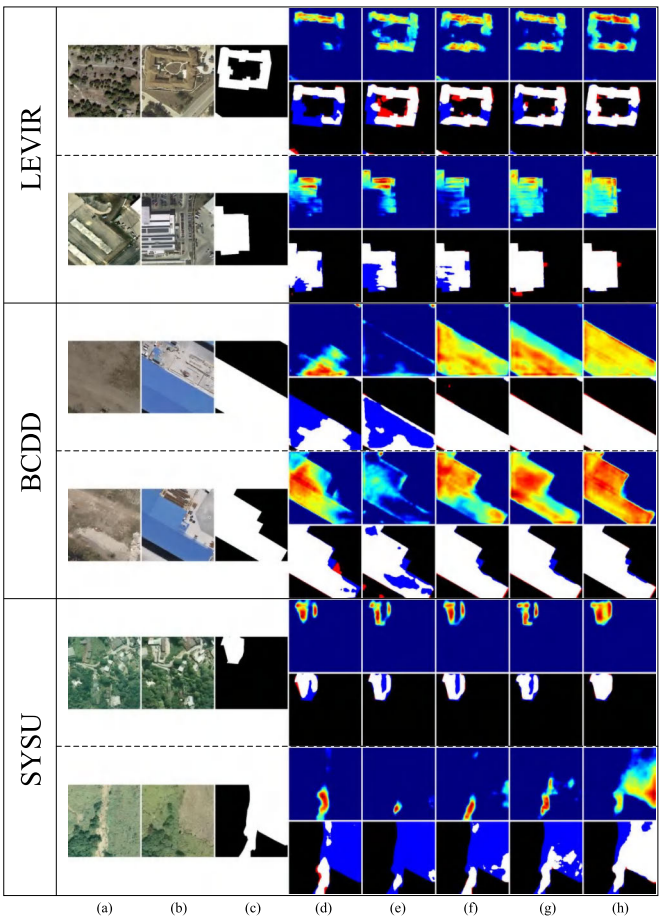

Fig. 12. Visual comparisons of the MDFM and the ablation baseline modules. (a) T1 images, (b) T2 images, (c) ground truth, (d) visualize and predicted results of concatenation, (e) visualize and predicted results of subtraction, (f) visualize and predicted results of DEM, (g) visualize and predicted results of TFIM, and (h) visualize and predicted results of MDFM.

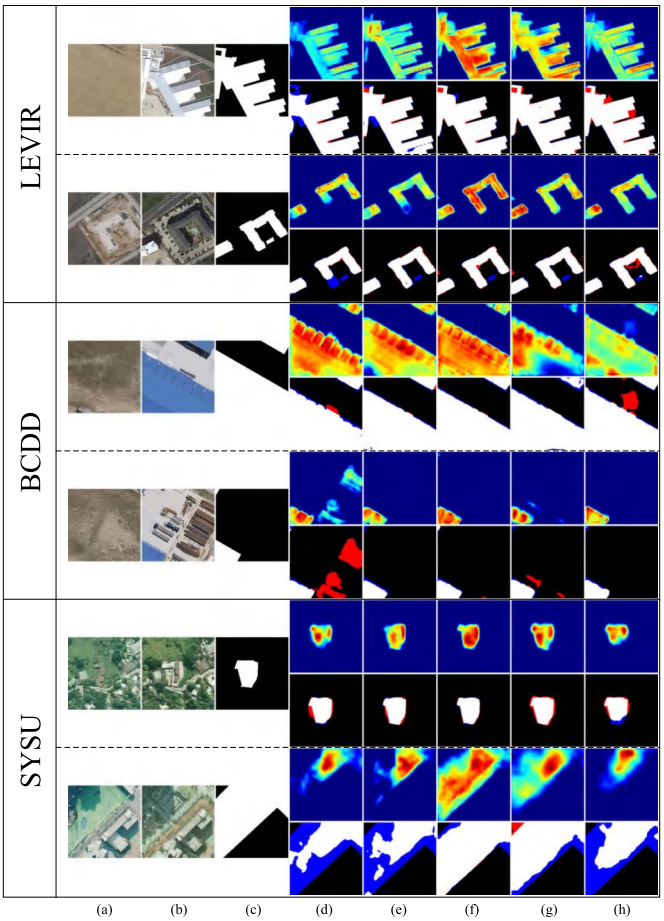

Fig. 13. Visual comparisons of the DAMs on three datasets. (a) T1 images, (b) T2 images, (c) ground truth, (d) visualize and predicted results of zero DAM, (e) visualize and predicted results of one DAM, (f) visualize and predicted results of two DAMs, (g) visualize and predicted results of three DAMs, and (h) visualize and predicted results of four DAMs.

gradually, which indicates that too many DAMs may overfit DGMA^2-Net in model training.

Fig. 13 visually shows the impacts of DAMs in DGMA^2-Net. It is noted that as more DAMs are employed in DGMA^2-Net (from zero to two DAMs), change areas are obviously detected with clear boundaries. This observation indicates that noises in the difference features are significantly reduced. When three or four DAMs are configured in DGMA^2-Net, changed areas are getting blurred and missed and false detections are being noticed.

To further demonstrate the advantages of DAMs, skip connections and guided refinement module (GRM) [21] are considered in the ablation experiments to compare with DAMs. Skip connections are used to replace the DAMs to aggregate bitemporal features and difference features from the MDFM, while the GRM conducts the same aggregation through concatenation and channel attention operations. Quantitative results are shown in Table V. It is observed that when DEMs are removed and skipconnections are used, the mertrics of the model downgrade. When GRM is used for feature aggregation, its recall in LEVIR-CD is slightly higher than those of DAM; however, its other mertrics are lower than that of DAM, showing that the DAMs are superior to skip connections and GRM in feature aggregation.

TABLE IV
ABLATION EXPERIMENT ON DAM. THE BEST VALUES ARE IN BOLD

Dataset	Number of DAM	Pre	Rec	F1	IoU	OA
LEVIR-CD	0	91.67	88.16	89.88	81.62	98.66
	1	**92.42**	88.66	90.50	82.65	99.05
	2	92.13	**90.31**	**91.21**	**83.84**	**99.29**
	3	90.67	89.92	90.29	82.30	99.01
	4	90.08	88.71	89.39	80.81	98.93
BCDD	0	93.13	91.44	92.28	85.66	99.38
	1	95.40	90.96	93.13	87.14	99.45
	2	**96.26**	**92.77**	**94.48**	**89.54**	**99.56**
	3	95.05	88.04	91.41	84.18	99.32
	4	89.84	91.49	90.66	82.92	98.23
SYSU-CD	0	81.67	82.19	81.93	69.39	91.45
	1	**85.30**	80.17	82.66	70.44	92.07
	2	84.32	83.24	**83.78**	**72.08**	**92.40**
	3	80.18	**83.52**	81.82	69.23	91.24
	4	84.98	76.25	80.38	67.19	91.22

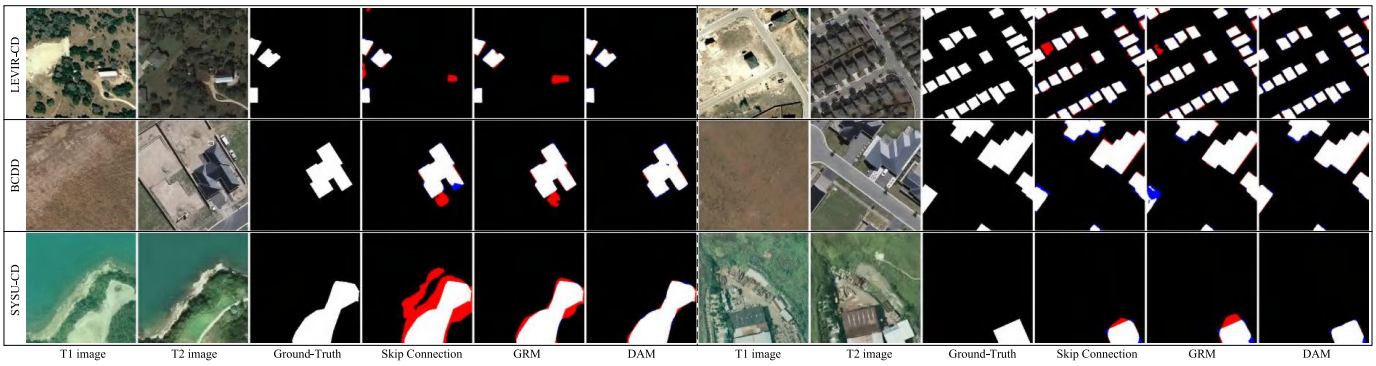

Fig. 14. Visual comparisons of DAMs, skip connections, and GRM on three datasets.

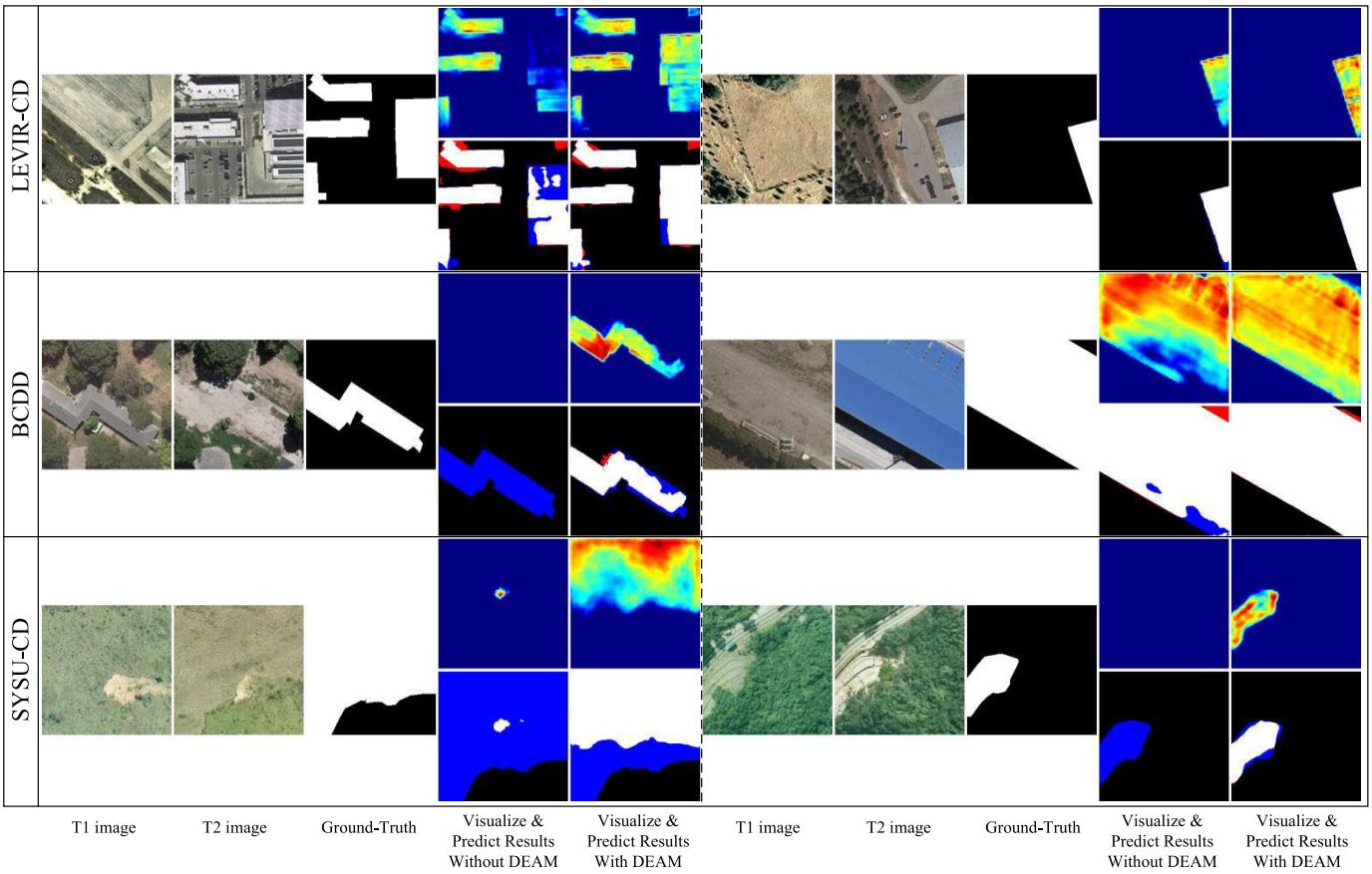

Fig. 15. Visual comparisons of the DGMA^2-Net with and without DEAMs.

Fig. 14 presents the predicted results of the DAMs, skip connections, and GRM. It can be found that when using skip connections and GRM, the model still has certain missed and false detections. When DAMs are used, these phenomena are reduced, confirming the importance of multilevel feature interaction in RSCD feature aggregation.

3) DEAM Ablation Experiment: The DEAM is developed to explore the interactions between bitemporal features and difference features, enhance the change areas, and refine difference features. Table VI lists the ablation experimental results of the DEAMs. When the DEAMs are removed, most metrics decrease. When DEAMs are added, the performance improves. Specifically, $F1$ has been improved by 0.56%, 1.38%, and 0.81%, while IoU has been improved by 0.92%, 2.44%, and 1.18% on the three datasets, respectively. This reflects that our DEAM can promote learning from bitemporal features and difference features.

The visual results of the DEAM ablation experiments are shown in Fig. 15. It is observed that the emergence of the DEAMs makes the model go deeper into changing areas, and reduces the false detections and missed detections. In addition, it shows that DEAM can effectively refine features, enhance the changing areas, and also yields excellent performance.

4) Module Efficiency Analysis: It is noted that in the absence of MDFM, we substitute subtraction. Table VII shows the efficiency analysis of the three proposed modules (MDFM,

TABLE V

COMPARISON OF DAM AND OTHER FEATURE AGGREGATION METHODS. THE BEST VALUES ARE IN BOLD

Dataset	Feature Aggregation	Pre	Rec	F1	IoU	OA
LEVIR-CD	Skip Connections	91.86	87.79	89.78	81.46	98.98
	GRM	90.58	**90.53**	90.55	82.74	99.04
	DAM	**92.13**	90.31	**91.21**	**83.84**	**99.14**
BCDD	Skip Connections	94.83	88.29	91.45	84.24	99.33
	GRM	96.04	90.87	93.38	87.58	99.47
	DAM	**96.26**	**92.77**	**94.48**	**89.54**	**99.56**
SYSU-CD	Skip Connections	83.58	78.39	80.90	67.92	91.27
	GRM	821.6	82.67	82.41	70.08	91.68
	DAM	**84.32**	**83.24**	**83.78**	**72.08**	**92.40**

TABLE VI

ABLATION EXPERIMENT ON DEAM. THE BEST VALUES ARE IN BOLD. O/W REPRESENTS THE DGMA^2-NET WITHOUT/WITH DEAM

Dataset	o/w DEAM	Pre	Rec	F1	IoU	OA
LEVIR-CD	without DEAM	90.49	**90.81**	90.65	82.89	99.05
	with DEAM	**92.13**	90.31	**91.21**	**83.84**	**99.14**
BCDD	without DEAM	94.69	91.56	93.10	87.10	99.45
	with DEAM	**96.26**	**92.77**	**94.48**	**89.54**	**99.56**
SYSU-CD	without DEAM	**87.89**	78.58	82.97	70.90	92.39
	with DEAM	84.32	**83.24**	**83.78**	**72.08**	**92.40**

TABLE VII

QUANTITATIVE ANALYSIS OF MODELS ON PARAMS, FLOPS, AS WELL AS $F1$, TRAIN TIME FOR ONE EPOCH, AND FPS ON LEVIR-CD DATASET. THE BEST VALUES ARE IN BOLD

Method	F1	Params(M)	Flops(G)	Epoch time(min)	FPS
DGMA^2-Net without MDFM	90.01	**24.55**	**10.35**	**1.9**	**35**
DGMA^2-Net without DAM	89.88	29.24	13.58	2.4	25
DGMA^2-Net without DEAM	90.65	36.67	18.10	2.8	25
DGMA^2-Net	**91.21**	37.10	18.10	2.9	22

DAM, and DEAM). It is noted that subtraction is used to conduct bitemporal feature fusion. It can be observed from the table that without MDFM, the model's Params, Flops, training efficiency, and FPS are better. In addition, it is observed that MDFM has the greatest impact on the efficiency of the model. For example, FPS dropped from 35 to 22, and the training time for one epoch increased from 1.9 to 2.9 min. Compared with MDFM, DAM, and DEAM have less impact on the efficiency of the model and can also improve the performance of the model. Notably, the multiscale convolution kernel and DAM consume more computing resources, and lightweight bitemporal feature fusion methods and feature aggregation methods need to be studied.

V. CONCLUSION

In this work, a difference-guided multiscale multilevel aggregation network, DGMA^2-Net, is developed for RSCD. DGMA^2-Net comprises three main components: MDFM, DAM, and DEAM. Initially, an ResNet-18 architecture is employed to extract multilevel features from the bitemporal images. The MDFMs are then used to fuse semantic information from bitemporal features into difference features through a series of convolutional operations with various kernel sizes. The MDFM is proven to be a new approach in fusing semantic information of bitemporal features in difference features. In addition, the DAMs are instrumental in effectively aggregating multilevel difference features and facilitating multilevel feature interactions. Finally, the DEAMs further enhance the detection of change areas through the interactions between difference features and bitemporal features.

The experimental results on three popular RSCD datasets (LEVIR-CD, BCDD, and SYSU-CD) demonstrate that DGMA^2-Net has exceptional performances in RSCD and surpasses the state-of-the-art baseline models. In addition, the ablation tests provide results that indicate that the MDFM, DAM, and DEAM are effective in fusing bitemporal features, refining difference features, and detecting change areas. The incorporation of MDFM, DAM, and DEAM within the framework of DGMA^2-Net introduces an innovative approach to RSCD that unequivocally outperforms the baseline RSCD models.

The DGMA^2-Net still requires more computing resources for RSCD, especially MDFM. With the continuous advancement of RSCD technologies, bitemporal feature fusion can be conducted by more lightweight and novel methods. Second, the multilevel feature aggregation in this article employed two DAMs. Therefore, future exploration could focus on investigating multilevel feature aggregation methods that only incorporate a single feature aggregation module. Furthermore, a new model is planned to incorporate semi-supervised, unsupervised, and self-supervised methods in RSCD since DGMA^2-Net requires a large amount of annotated labels.

REFERENCES

[1] L. Gueguen and R. Hamid, "Toward a generalizable image representation for large-scale change detection: Application to generic damage analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3378–3387, Jun. 2016.

[2] W. Zhang, L. Shen, and W. Qiao, "Building damage detection in VHR satellite images via multi-scale scene change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2021, pp. 8570–8573.

[3] A. Reigber, M. Jäger, and E. Krogager, "Polarimetric SAR change detection in multiple frequency bands for environmental monitoring in Arctic regions," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 5702–5705.

[4] S. Shi, Y. Zhong, J. Zhao, P. Lv, Y. Liu, and L. Zhang, "Land-use/land-cover change detection based on class-prior object-oriented conditional random field framework for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5600116.

[5] M. Che and P. Gamba, "Intra-urban change analysis using Sentinel-1 and nighttime light data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 4, pp. 1134–1142, Apr. 2019.

[6] F. Bovolo and L. Bruzzone, "A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 1, pp. 218–236, Jan. 2007.

[7] G. F. Byrne, P. F. Crapper, and K. K. Mayo, "Monitoring land-cover change by principal component analysis of multitemporal Landsat data," *Remote Sens. Environ.*, vol. 10, no. 3, pp. 175–184, Nov. 1980.

[8] B. Du, L. Ru, C. Wu, and L. Zhang, "Unsupervised deep slow feature analysis for change detection in multi-temporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9976–9992, Dec. 2019.

[9] S. Cui and C. Luo, "Feature-based non-parametric estimation of Kullback–Leibler divergence for SAR image change detection," *Remote Sens. Lett.*, vol. 7, no. 11, pp. 1102–1111, Nov. 2016.

[10] J. Im, J. R. Jensen, and J. A. Tullis, "Object-based change detection using correlation image analysis and image segmentation," *Int. J. Remote Sens.*, vol. 29, no. 2, pp. 399–423, Jan. 2008.

[11] X. Wei et al., "Fine-grained image analysis with deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8927–8948, Dec. 2022.

[12] L. Zheng, Y. Yang, and Q. Tian, "SIFT meets CNN: A decade survey of instance retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1224–1244, May 2018.

[13] M. E. Paoletti, S. Moreno-Álvarez, Y. Xue, J. M. Haut, and A. Plaza, "AAtt-CNN: Automatic attention-based convolutional neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5700314.

[14] X. Guo, B. Hou, B. Ren, Z. Ren, and L. Jiao, "Network pruning for remote sensing images classification based on interpretable CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5605615.

[15] J. Pang, C. Li, J. Shi, Z. Xu, and H. Feng, "$\mathcal{R}^2$-CNN: Fast tiny object detection in large-scale remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5512–5524, Aug. 2019.

[16] M. Li, X. Zhao, J. Li, and L. Nan, "ComNet: Combinational neural network for object detection in UAV-borne thermal images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6662–6673, Aug. 2021.

[17] R. Guan, M. Wang, L. Bruzzone, H. Zhao, and C. Yang, "Lightweight attention network for very high-resolution image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4403514.

[18] H. Jung, H.-S. Choi, and M. Kang, "Boundary enhancement semantic segmentation for building extraction from remote sensed image," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5215512.

[19] Z. Li, C. Tang, L. Wang, and A. Y. Zomaya, "Remote sensing change detection via temporal feature interaction and guided refinement," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5628711.

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

[21] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5604816.

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

[23] 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 Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1845–1849, Oct. 2017.

[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 Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 266–270, Feb. 2019.

[25] 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 Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 3, pp. 545–559, Dec. 2016.

[26] X. Li, L. Yan, Y. Zhang, and N. Mo, "SDMNet: A deep-supervised dual discriminative metric network for change detection in high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[27] A. Eftekhari, F. Samadzadegan, and F. Dadrass Javan, "Building change detection using the parallel spatial-channel attention block and edge-guided deep network," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 117, Mar. 2023, Art. no. 103180.

[28] W. Zhang, Q. Zhang, H. Ning, and X. Lu, "Cascaded attention-induced difference representation learning for multispectral change detection," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 121, Jul. 2023, Art. no. 103366.

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

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

[31] 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 Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4409818.

[32] Z. Li et al., "Lightweight remote sensing change detection with progressive feature aggregation and supervised attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5602812.

[33] W. Gao, Y. Sun, X. Han, Y. Zhang, L. Zhang, and Y. Hu, "AMIO-net: An attention-based multiscale input–output network for building change detection in high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 2079–2093, 2023.

[34] Y. Lei, D. Peng, P. Zhang, Q. Ke, and H. Li, "Hierarchical paired channel fusion network for street scene change detection," *IEEE Trans. Image Process.*, vol. 30, pp. 55–67, 2021.

[35] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected Siamese network for change detection of VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021.

[36] Y. Zhao, P. Chen, Z. Chen, Y. Bai, Z. Zhao, and X. Yang, "A triple-stream network with cross-stage feature fusion for high-resolution image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5600417.

[37] J. Zhang et al., "AERNet: An attention-guided edge refinement network and a dataset for remote sensing building change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5617116.

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

[39] D. Song, Y. Dong, and X. Li, "Context and difference enhancement network for change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9457–9467, 2022.

[40] 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 Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2021.

[41] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

[42] A. Raza, H. Huo, and T. Fang, "EUNet-CD: Efficient UNet++ for change detection of very high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[43] X. Zhang et al., "DifUnet++: A satellite images change detection network based on Unet++ and differential pyramid," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[44] Q. Li, R. Zhong, X. Du, and Y. Du, "TransUNetCD: A hybrid transformer network for change detection in optical remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622519.

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

[46] J. Zhang, M. Xing, G.-C. Sun, and X. Shi, "Vehicle trace detection in two-pass SAR coherent change detection images with spatial feature enhanced Unet and adaptive augmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5232415.

[47] Y. Sun, L. Lei, X. Tan, D. Guan, J. Wu, and G. Kuang, "Structured graph based image regression for unsupervised multimodal change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 185, pp. 16–31, Mar. 2022.

[48] Y. Sun, L. Lei, D. Guan, M. Li, and G. Kuang, "Sparse-constrained adaptive structure consistency-based unsupervised image regression for heterogeneous remote-sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4405814.

[49] E. P. Crist, "A TM tasseled cap equivalent transformation for reflectance factor data," *Remote Sens. Environ.*, vol. 17, no. 3, pp. 301–306, Jun. 1985.

[50] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and k-means clustering," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 772–776, Oct. 2009.

[51] U. H. Atasever and M. A. Gunen, "Change detection approach for SAR imagery based on arc-tangential difference image and k-means++," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[52] T. Xiao, Y. Wan, J. Chen, W. Shi, J. Qin, and D. Li, "Multiresolution-based rough fuzzy possibilistic C-means clustering method for land cover change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 570–580, 2023.
[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[54] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
[55] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2021.
[56] X. Zhao, K. Zhao, S. Li, and X. Wang, "GeSANet: Geospatial-awareness network for VHR remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5402814.
[57] Y. Feng, H. Xu, J. Jiang, H. Liu, and J. Zheng, "ICIF-Net: Intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4410213.
[58] S. Fang, K. Li, and Z. Li, "Changer: Feature interaction is what you need for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5610111.

Zilu Ying received the B.S., M.S., and Ph.D. degrees in electrical information engineering from Beihang University, Beijing, China, in 1985, 1988, and 2009, respectively.

He is currently a Full Professor with Wuyi University, Jiangmen, China. His research interests include biometric extraction and pattern recognition.

Dr. Ying is an Executive Director of the Guangdong Society of Image and Graphics and a member of the Signal Processing Branch, Chinese Institute of Electronics, Beijing.

Zijun Tan received the B.S. degree from Wuyi University, Jiangmen, China, in 2021, where he is currently pursuing the master's degree with the Department of Intelligence Manufacturing.

His research interests include semantic segmentation, change detection, and pattern recognition.

Yikui Zhai (Senior Member, IEEE) received the Ph.D. degree in signal and information processing from Beihang University, Beijing, China, in 2013.

Since October 2007, he has been working with Wuyi University, Jiangmen, China, where he is currently a Full Professor and he has been an Associate Dean of the School of Electronics and Information Engineering, since 2021. He has been a Visiting Scholar with the Department of Computer Science, Università degli Studi di Milano, Milan, Italy, from June 2016 to June 2017, August 2023 and January 2024.

His research interests include image processing, deep learning, optical character recognition, object detection, unmanned aerial vehicle (UAV) change detection, and self-supervised learning.

Xudong Jia received the B.S. and M.S. degrees from Beijing Jiaotong University, Beijing, China, in 1983 and 1986, respectively, the second M.S. degree from the University of Toronto, Toronto, ON, Canada, in 1992, and the Ph.D. degree from the Georgia Institute of Technology, Atlanta, GA, USA, in 1996.

He is currently a Professor and the Associate Dean of the College of Engineering and Computer Science, California State University at Northridge (CSUN), Los Angeles, CA, USA. His research interests include intelligent transportation systems (ITS) standards, geographic information system (GIS) applications in transportation, traffic safety, transportation information systems, travel demand management, and air quality.

Dr. Jia is an Associate Editor of the IEEE Intelligent Transportation Systems Society and IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS.

Wenba Li (Student Member, IEEE) received the B.S. degree from Guangdong University of Technology, Guangzhou, China, in 2020. He is currently pursuing the master's degree with the Department of Intelligence Manufacturing, Wuyi University, Jiangmen, China.

His research interests include image classification, change detection, and pattern recognition.

Junying Zeng (Member, IEEE) received the Ph.D. degree in physical electronics from Beijing University of Posts and Telecommunications, Beijing, China, in 2008.

He is currently a Professor with Wuyi University, Jiangmen, Guangdong, China. His research interests include intelligent signal processing and pattern recognition.

Angelo Genovese (Senior Member, IEEE) received the Ph.D. degree in computer science from the Università degli Studi di Milano, Crema, Italy, in 2014.

He has been a Post-Doctoral Research Fellow in computer science with the Università degli Studi di Milano, Milan, since 2014. He was a Visiting Researcher with the University of Toronto, Toronto, ON, Canada. His original results have been published in over 30 articles in international journals, proceedings of international conferences, books, and book chapters. His research interests include signal and image processing, 3-D reconstruction, computational intelligence technologies for biometric systems, industrial and environmental monitoring systems, and design methodologies and algorithms for self-adapting systems.

Vincenzo Piuri (Fellow, IEEE) received the M.S. and Ph.D. degrees in computer engineering from the Politecnico di Milano, Milan, Italy, in 1984 and 1988, respectively.

He was the Department Chair with the University of Milan, Milan, from 2007 to 2012, where he has been a Full Professor since 2000. He was an Associate Professor with the Politecnico di Milano, from 1992 to 2000, a Visiting Professor with The University of Texas at Austin, Austin, TX, USA, from 1996 to 1999, and a Visiting Researcher with George Mason University, Fairfax, VA, USA, from 2012 to 2016. He founded a start-up company, Sensure srl, Bergamo, Italy, in the area of intelligent systems for industrial applications (leading it from 2007 to 2010) and was active in industrial research projects with several companies. His main research and industrial application interests are intelligent systems, computational intelligence, pattern analysis and recognition, machine learning, signal and image processing, biometrics, intelligent measurement systems, industrial applications, distributed processing systems, the Internet-of-Things, cloud computing, fault tolerance, application-specific digital processing architectures, and arithmetic architectures.

Dr. Piuri is an ACM Fellow.

Fabio Scotti (Senior Member, IEEE) received the Ph.D. degree in computer engineering from the Politecnico di Milano, Milan, Italy, in 2003.

He was an Assistant Professor with the Department of Information Technologies, Università degli Studi di Milano, Milan, from 2002 to 2015, where he was an Associate Professor with the Department of Computer Science from 2015 to 2020. He has been a Full Professor with the Università degli Studi di Milano since 2020. His original results have been published in over 150 articles in international journals, proceedings of international conferences, books, book chapters, and patents. His research interests include biometric systems, machine learning and computational intelligence, signal and image processing, theory and applications of neural networks, 3-D reconstruction, industrial applications, intelligent measurement systems, and high-level system design.

Dr. Scotti is an Associate Editor of IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS and the IEEE OPEN JOURNAL OF SIGNAL PROCESSING. He is serving as a Book Editor (Area Editor, section Less-Constrained Biometrics) of the *Encyclopedia of Cryptography, Security, and Privacy* (Third Edition, Springer). He has been an Associate Editor of IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY and *Soft Computing* (Springer) and a Guest Co-Editor of IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT.