IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 23, 2022, accepted April 18, 2022, date of publication April 26, 2022, date of current version May 3, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3170704

Robust Change Detection Using Channel-Wise Co-Attention-Based Siamese Network With Contrastive Loss Function

EUNJEONG CHOI AND JEONGTAE KIM
Department of Electronic and Electrical Engineering, Ewha Womans University, Seoul 03760, South Korea
Corresponding author: Jeongtae Kim (jtkim@ewha.ac.kr)

This work was supported by the ATI Company in Incheon, Republic of Korea.

ABSTRACT Change detection methods aim to identify significantly changed areas in co-registered bitemporal images taken of the same area. Since not only do bitemporal images usually have different environmental conditions (*i.e.,* different weather conditions, noises, and seasonal changes) but also changes irrelevant to the purpose of change detection (*e.g.,* road changes when detecting building change), which should not be detected as changed areas, change detection methods often suffer from the problem of pseudo-change detection. To alleviate this problem, we propose an encoder-decoder-based Siamese network with a channel-wise co-attention module that considers the channel-wise correlations between a feature map in one image and all feature maps in the other image. By comparing the feature map in one image with the revised feature map in the other image considering the correlations, we are able to reduce the differences between the feature maps when pseudo-changes exist, thereby rendering the proposed method more robust to pseudo-changes. In addition, we apply a contrastive loss function that encourages the pairs of feature maps corresponding to unchanged regions to be similar, which can help improve the performance of change detection. We verified the performance of the proposed method through experiments using datasets such as the change detection dataset (CDD) and building change detection dataset (BCDD). In the experiment, the proposed method achieved significantly improved performance compared with existing methods in terms of recall, precision, f1-score, and overall accuracy.

INDEX TERMS Attention, change detection, co-attention, deep learning, remote sensing, Siamese network.

I. INTRODUCTION

Change detection (CD) is the task of identifying changed areas in two co-registered images of the same location acquired at different times [1]. CD methods usually assign a binary label to each pixel in a *target* image (also called a T_1 image) to indicate whether or not the pixel belongs to the changed area from the *reference* image (also called a T_0 image) [2]–[5]. Although identifying changed pixels based on the intensity values may seem straightforward, CD is a challenging task due to the existence of pseudo-changes, which should not be detected as genuine changes even though the intensity values of the corresponding pixels are significantly different. For example, pseudo-changes may be generated due to environmental changes in two images as a result of illumination changes or seasonal changes, as shown in Fig.1. Fig.1 shows example images from the change detection dataset (CDD) [6] and Figs.1(a) and (b) show bitemporal image pairs with illumination change and seasonal change, respectively. Although the two images are very different in their pixel values, CD should determine that the areas in the two images have not changed if the difference is caused by a pseudo-change [7]. Even more challenging pseudo-changes are application-specific pseudo-changes that can be pseudo or genuine changes depending on the purpose of the CD. For example, in the field of urbanization monitoring, changes related to buildings are genuine changes while changes related to trees are pseudo-changes [2], [3]. On the contrary, in the field of deforestation monitoring, changes related to trees are genuine changes [8].

Many CD methods have been studied using various image processing methods [9]–[16]. Recently, with the successful application of deep learning to computer vision and remote sensing [7], [17]–[21], deep learning-based CD methods

The associate editor coordinating the review of this manuscript and approving it for publication was Manuel Rosa-Zurera.

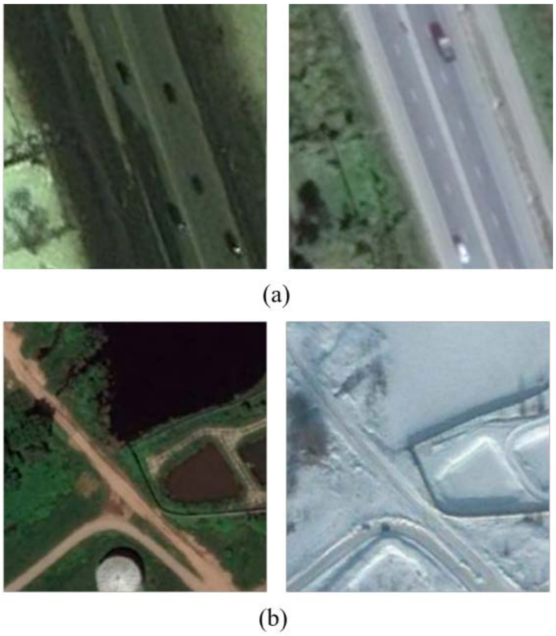

FIGURE 1. Illustrations of bitemporal image pairs in the CDD dataset [6]: (a) image pair with illumination change and (b) image pair with environmental change.

have attracted much attention [7], [22]–[29]. Some methods apply deep learning networks to extract feature maps from two images and then calculate the distance between the feature maps to generate a change map [7], [22], [23]. Other methods combine feature maps extracted from two images and then decode the combined feature maps to generate a change map [24], [26]–[28].

Many deep learning-based CD methods use a Siamese network that contains two identical networks that share weights [7], [22], [23], [26]–[28]. Since Siamese network-based CD methods identify changed areas based on the difference between two images, these methods often suffer from pseudo-changes that may cause large differences in intensity values. To alleviate the problem of pseudo-change detection, some CD methods have applied attention modules that can help obtain more discriminant feature representations by capturing feature dependencies [7], [22], [23], [27], [30]–[32]. Some methods have applied self-attention modules that consider feature dependencies within a single image to better distinguish between changed areas and unchanged areas [7], [30]–[32]. However, for improving robustness to pseudo-changes, we believe that feature dependencies between bitemporal images should be considered rather than dependencies within a single image. Although several CD methods have applied co-attention modules that consider spatial-wise feature dependencies between bitemporal images, they focus more on reducing errors caused by misregistration than reducing pseudo-change detection [22], [27].

Inspired by the above-mentioned attention modules, we propose a channel-wise co-attention-based Siamese network for CD, which we expect to be more robust to pseudo-changes than existing methods. The channel-wise co-attention module considers channel-wise feature dependencies between bitemporal images. The idea behind the proposed method is that if an area belongs to a pseudo-change area, there may exist a similar feature map in another image even though the map may be from a different channel. Although a direct comparison between feature maps can result in large differences, if we find a similar feature map in the other image and compute the differences with the similar feature map, the differences will be small. Based on this idea, we believe that the proposed method can help reduce the detection of pseudo-changes. In addition, to improve the performance of CD, we apply a contrastive loss function that encourages the distance between features from unchanged regions to be small and the distance between features from changed regions to be larger than a specific margin.

In this paper, we quantitatively demonstrate that the proposed method can improve the performance of CD. The proposed method shows superior performance compared with existing methods for two open datasets: the change detection dataset (CDD) [6] and building change detection dataset (BCDD) [33].

The remainder of this paper is organized as follows. In Section 2, we review related studies. We also explain the proposed method in detail in Section 3. The experimental results and conclusions are presented in Sections 4 and 5.

II. RELATED WORKS

A. SIAMESE NETWORK-BASED CHANGE DETECTION

A Siamese network is a neural network that contains two sub-networks [19] and uses two images for input. The network extracts features from the two images in parallel using each sub-network and then considers the difference between the extracted features for image comparison [19]. By sharing the weights of the sub-networks, a Siamese network can identify whether similar features exist, thereby comparing the two images more effectively.

Although several CD methods such as the fully convolutional early fusion network (FC-EF) [24] and the boundary-aware attentive network (BA2Net) [25] are based on U-Net [18], recently, much research has been focused on Siamese network-based CD methods. Some methods extract feature maps from bitemporal images using a Siamese network and then calculate the distance between the feature maps to generate a change map. These methods include a dual attentive fully convolutional Siamese network (DASNet) [7], a spatial-temporal attention network (STANet) [22], and a deeply supervised attention metric network (DSAMNet) [23]. DASNet [7] uses self-attention modules to obtain more discriminant feature representations and attempts to reduce the detection of pseudo-change. In STANet [22], a basic spatial-temporal attention module (BAM) and a pyramid spatial-temporal attention module (PAM) are used to obtain illumination-invariant and misregistration-robust features. DSAMNet [23] uses a combined attention module to

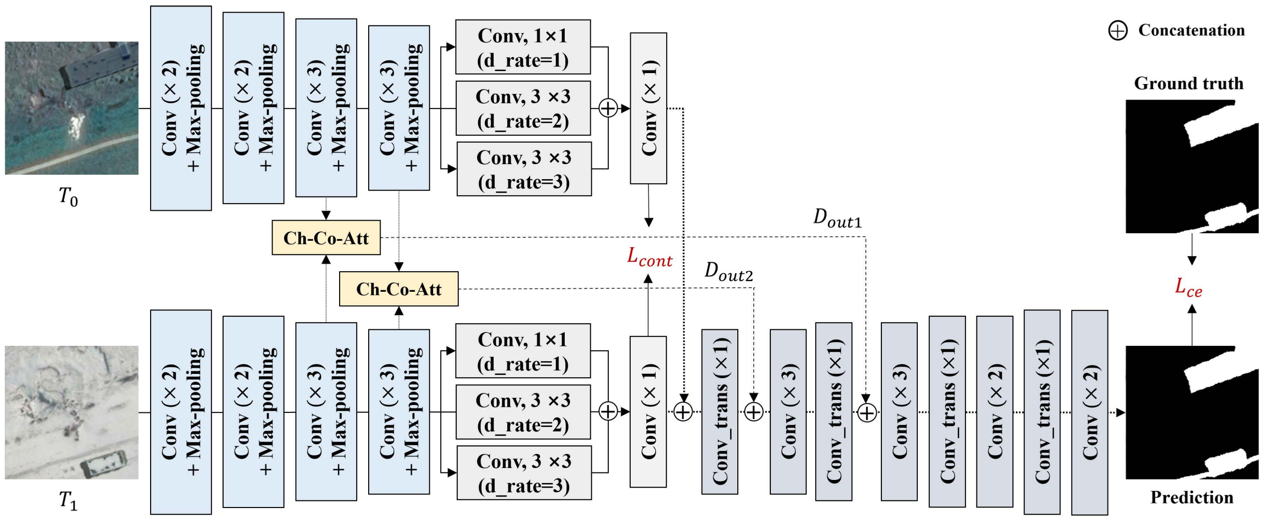

FIGURE 2. Overview of the proposed method: Ch-Co-Att, Conv, and Conv_trans represent the channel-wise co-attention module, convolution layer, and transposed convolution layer, respectively. All convolution and transposed convolution layers are followed by batch normalization and ReLU layers.

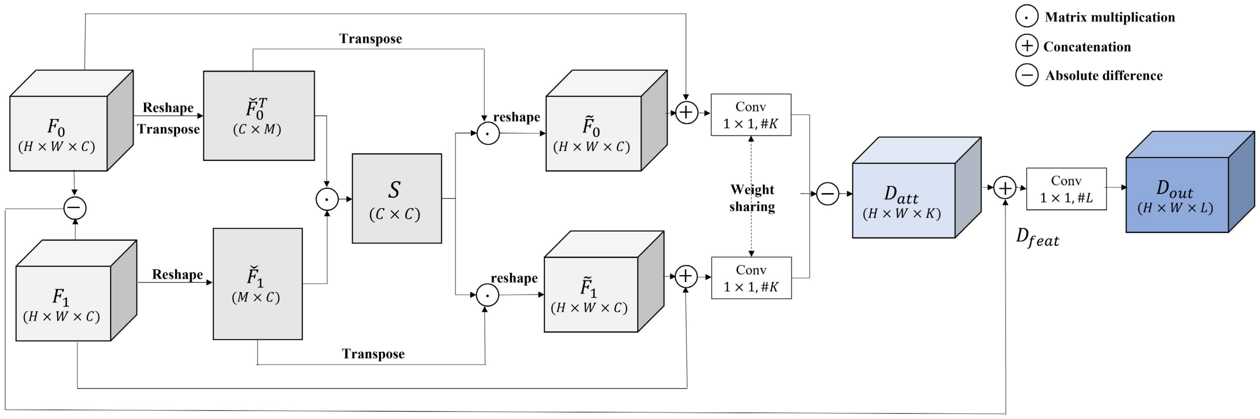
Matrix multiplication
Concatenation
Absolute difference
Transpose
F_0 $(H \times W \times C)$
Reshape
Transpose
$\breve{F}_0^T$ $(C \times M)$
reshape
$\tilde{F}_0$ $(H \times W \times C)$
Conv $1 \times 1, \#K$
Weight sharing
S $(C \times C)$
D_{att} $(H \times W \times K)$
Conv $1 \times 1, \#L$
D_{out} $(H \times W \times L)$
D_{feat}
F_1 $(H \times W \times C)$
Reshape
$\breve{F}_1$ $(M \times C)$
reshape
$\tilde{F}_1$ $(H \times W \times C)$
Conv $1 \times 1, \#K$
Transpose

FIGURE 3. Overview of the proposed channel-wise co-attention module.

recognize pseudo-changes. Other methods combine feature maps extracted from bitemporal images using a Siamese network and then decode the combined feature maps to generate a change map. These methods include a fully convolutional Siamese network with concatenation skip connections (FC-Siam-Concat) [24], a fully convolutional Siamese network with difference skip connections (FC-Siam-Diff) [24], and a pyramid feature-based attention-guided Siamese network (PGA-SiamNet) [27].

B. SELF-ATTENTION MODULE

Recently, some CD methods have applied self-attention modules to improve CD performance [7], [22], [30]–[32]. The self-attention module considers feature dependencies within a single image. For a feature vector at one position, the self-attention module calculates the correlations between the feature vector and all other feature vectors in the same image to generate a refined feature vector, which is computed by the linear combination of all other feature vectors using the calculated correlations [7], [21], [22], [30]–[32], [34], [35]. Since the refined feature vector may reflect features of objects belonging to the same category but with different appearances, using a refined feature vector may render CD methods robust to pseudo-changes. However, since the features of one image are compared with the features of the other image in CD tasks, it is more effective to generate refined feature vectors using feature vectors of the other image to reduce pseudo-change detection. Based on this idea, studies on spatial-wise co-attention modules have been conducted [22], [27].

C. SPATIAL-WISE CO-ATTENTION MODULE

As mentioned above, spatial-wise co-attention modules that consider spatial-wise correlations between bitemporal images may help reduce the detection of pseudo-changes. In addition, spatial-wise co-attention modules can be helpful for reducing errors due to misregistration because a feature vector in an image is compared with a similar feature vector in the other image regardless of position [22], [27]. However, we think that the performance improvement provided by a spatial-wise co-attention module is limited because misregistration is minimal for most CD methods since co-registered bitemporal images are used. In terms of pseudo-change detection, because the module computes the refined feature vector using similar feature vectors in the other image, if the same structure in the other image generates different feature vectors due to pseudo-changes, these pseudo-changes may be detected as a genuine change since the feature vectors are different. To alleviate this problem, we believe that the inclusion of the channel-wise co-attention module that considers channel-wise correlations between bitemporal images is beneficial since similar features may be present in different channels due to pseudo-changes. Based on this observation, we propose a novel CD method with channel-wise co-attention module, which is explained in Section 3.

III. PROPOSED METHOD

In this section, we present our channel-wise co-attention-based CD method. We first explain the overall network architecture and a novel channel-wise co-attention module for CD in detail. Finally, we describe total loss functions for the proposed method.

A. NETWORK ARCHITECTURE

We propose an encoder-decoder-based Siamese network with channel-wise co-attention modules. Fig.2 shows the structure of the proposed network, which consists of an encoder network with an atrous spatial pyramid pooling (ASPP) module [36] to extract high-level features without compromising spatial resolution, the channel-wise co-attention module, and a decoder network. As shown in Fig.2, the encoder extracts feature maps from two images in parallel using the same convolution layers and merges the extracted feature maps to use them as input to the decoder. Then, the decoder determines whether or not a pixel belongs to a changed area using the convolution layers with the merged feature maps and feature maps from the co-attention module, which is explained later. As one may intuitively expect, the feature maps of the changed areas in the two images may have large differences, while those of the unchanged areas show only moderate differences as long as there are no pseudo-changes. However, if pseudo-changes exist, the corresponding feature maps can be significantly different because the intensity values for the areas can be significantly different.

To reduce the problem of pseudo-change detection, we apply channel-wise co-attention modules to the outputs of the 3rd and 4th convolutional blocks of the encoder (*i.e.,* feature maps right before max-pooling layer) and transfer the output of each co-attention module to the corresponding convolutional block of the decoder in the form of skip-connections. Given bitemporal image pair T_0 and T_1, we denote the extracted feature maps from T_0 and T_1 with $F_0 \in \mathbb{R}^{H\times W\times C}$ and $F_1 \in \mathbb{R}^{H\times W\times C}$, respectively, where H is the height, W is the width, and C is the number of channels. The channel-wise co-attention module computes the correlations between the i^{th} feature map in one image and all the feature maps in the other image and then uses the correlations as weights for computing a linear combination of the feature maps in the other image. We use the linear combination as the generated feature map, expecting that the generated feature map is similar to the i^{th} feature map in one image even though there exist pseudo-changes. The intuition behind this expectation is that there may be similar feature maps in the other image in different channels since the different characteristics of the two images may show similar shapes under pseudo-changes. For example, if two images are acquired under different lighting conditions, features from different color filters may have similar shapes.

The aforementioned operation of the channel-wise co-attention module is implemented as shown in Fig.3. First, we compute the affinity matrix $S \in \mathbb{R}^{C\times C}$ between F_0 and F_1 using the cosine similarity as follows:

$$s_{ij} = \frac{\check{F}_0^{(j)} \cdot \check{F}_1^{(i)}}{\|\check{F}_0^{(j)}\|\|\check{F}_1^{(i)}\|} \tag{1}$$

where s_{ij} is the j^{th} element in the i^{th} column of affinity matrix $S \in \mathbb{R}^{C\times C}$ and represents the degree of similarity between the j^{th} feature map of F_0 and the i^{th} feature map of F_1. $\check{F}_0 \in \mathbb{R}^{M\times C}$ and $\check{F}_1 \in \mathbb{R}^{M\times C}$ are reshaped feature maps from F_0 and F_1, respectively, M is the multiplication of H and W, $\check{F}_0^{(j)}$ denotes the j^{th} column of $\check{F}_0$, and $\check{F}_1^{(i)}$ denotes the i^{th} column of $\check{F}_1$.

Next, given two reshaped feature maps $\check{F}_0$ and $\check{F}_1$, the channel-wise co-attention module generates modified feature maps considering the channel-wise correlations with feature maps from the other image as follows:

$$\begin{aligned}\hat{F}_0 &= \check{F}_0 S = [\hat{F}_0^{(1)}, \hat{F}_0^{(2)}, \ldots \hat{F}_0^{(i)} \ldots \hat{F}_0^{(C)}] \in \mathbb{R}^{M\times C}\\ \hat{F}_0^{(i)} &= \frac{1}{C}\sum_{j=1}^{C} \check{F}_0^{(j)} s_{ij} \in \mathbb{R}^{M},\end{aligned} \tag{2}$$

where $\hat{F}_0$ represents the revised feature maps from $\check{F}_0$. $\hat{F}_0^{(i)}$ indicates the i^{th} column of $\hat{F}_0$ and reflects the correlations between the i^{th} feature map of F_1 and all feature maps of F_0. Similarly, the revised feature maps $\hat{F}_1$ can be computed by $\hat{F}_1 = S\check{F}_1^T$.

For a feature map in one image, if similar feature maps exist in the other image in different channels, the difference between the same channel feature maps may be large, but the difference between the feature map in one image and similar feature maps in the other image can be small enough to avoid detection of the pseudo-changes. Based on this, we consider the difference between the feature maps in one image and the feature maps obtained from the channel-wise co-attention as follows:

$$D_{att} = \text{abs}(f_g([F_0; \tilde{F}_0]) - f_g([\tilde{F}_1; F_1])) \in \mathbb{R}^{H\times W\times K} \tag{3}$$

where $\tilde{F_0} \in \mathbb{R}^{H\times W\times C}$ and $\tilde{F_1} \in \mathbb{R}^{H\times W\times C}$ are reshaped feature maps from $\hat{F}_0$ and $\hat{F}_1$, respectively, [;] denotes the concatenate operation, and f_g is a 1 × 1 convolution layer in which the number of filters is K. In addition to D_{att}, we also use the same channel-wise difference between the feature maps F_0 and F_1 to prevent the proposed method from missing the detection of changed areas as follows:

$$D_{feat} = \text{abs}(F_0 - F_1) \in \mathbb{R}^{H\times W\times C}. \tag{4}$$

We combine the difference maps obtained from Equation (3) and (4) to reduce detection of pseudo-changes without compromising the performance of detecting genuine change. The optimal combination is determined during training using trainable weights as follows:

$$D_{out} = f_h([D_{att}; D_{feat}]) \in \mathbb{R}^{H\times W\times L} \tag{5}$$

where f_h is a 1 × 1 convolution layer in which the number of filters is L. The combined difference maps D_{out} are transferred to the corresponding decoder layer in the form of skip connections.

B. LOSS FUNCTION

We also incorporate a loss function to reduce detection of pseudo-changes. In addition to the usual cross-entropy loss function for labeled data, we add a contrastive loss function that encourages the difference between features from the unchanged areas of the two images to be small while enforcing features from the changed areas to be larger than a specific margin. The contrastive loss function was shown to be effective in improving the performance of CD in previous investigations [7], [22]. We compute the contrastive loss function using outputs of the encoder as

$$L_{cont} = \sum_{i,j} w(1-y_{ij})d_{ij}^2 + (1-w)y_{ij}[\max(m-d_{ij},0)]^2, \tag{6}$$

where d_{ij} is the distance between the feature vectors of F_0 and F_1 at position (i,j) and m is the margin for changed feature pairs. w is used to balance the weights of the two terms in Equation (6), and y_{ij} is a label at position (i,j). We set the label in the changed area to 1 and the label in the unchanged area to 0. Therefore, the first term is zero in the changed areas while the second term is zero in the unchanged areas. By minimizing the loss function, the distance between the features from the unchanged areas should be close to zero because the first term of Equation (6), wd_{ij}^2, is minimized. On the contrary, the distance between the features from the changed areas should be larger than margin m because the second term of Equation (6), $(1-w)[\max(m-d_{ij},0)]^2$, is minimized.

The total loss function of the proposed method is defined as

$$L_{total} = \lambda L_{cont} + (1-\lambda)L_{ce}, \tag{7}$$

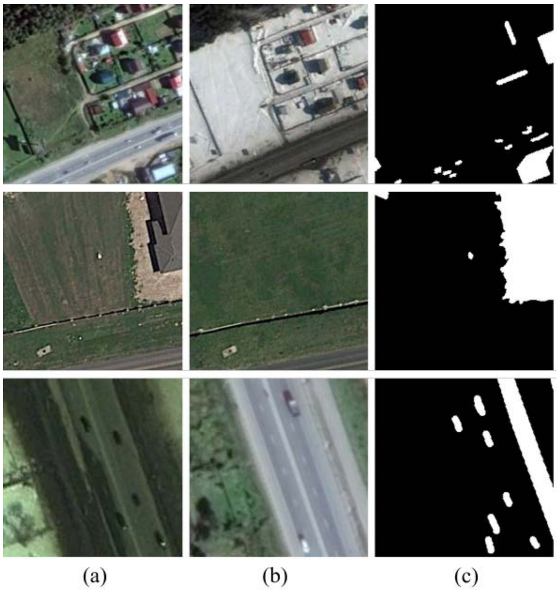

FIGURE 4. Illustrations of the CDD dataset [6]: (a) T_0 image, (b) T_1 image, and (c) ground truth.

where L_{cont} is the contrastive loss function, L_{ce} is the weighted cross-entropy loss function between the prediction and ground truth, and λ is the weight between the two losses.

IV. EXPERIMENT

To verify the effectiveness of the proposed method, we compared the performance of the proposed method with that of conventional CD methods such as FC-EF [24], FC-Siam-Conc [24], FC-Siam-Diff [24], DASNet [7] that uses self-attention modules, and STANet [22] that uses both self-attention module and spatial-wise co-attention module. We conducted experiments involving the detection of changes in the well-known CDD [6] and BCDD [33] datasets.

A. DATASETS

1) CDD DATASET

The CDD dataset is a remote sensing change detection dataset that is open to the public [6]. The dataset contains 11 full-size image pairs of season-varying images, of which 7 image pairs are 2, 700 × 4, 725 pixels and 4 image pairs are 1, 000 × 1, 900 pixels [6]. The spatial resolutions of the images in the dataset are between 3 cm to 100 cm per pixel. In [6], the original image pairs are cropped into images that are 256 × 256 pixels to generate a cropped dataset that contains 10,000 images for training, 3,000 images for test, and 3,000 images for validation. We used the cropped datasets for this investigation. Fig.4 shows example images of the cropped CDD dataset, Figs.4(a) and (b) show bitemporal image pairs, and Fig.4(c) shows the ground truth images.

2) BCDD DATASET

The BCDD dataset covers an area that was rebuilt after the occurrence of a 6.3-magnitude earthquake in

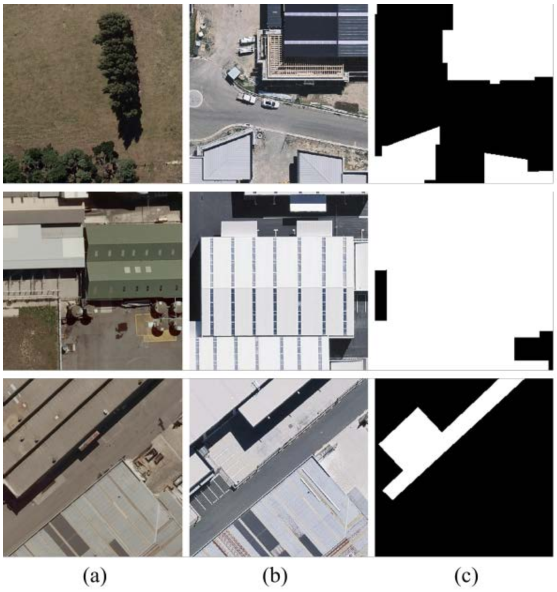

FIGURE 5. Illustrations of the BCDD dataset [33]: (a) T_0 image, (b) T_1 image, and (c) ground truth.

February 2011 [33]. The main purpose of the dataset is to detect changes related to buildings before and after the earthquake. Because the dataset contains only one image pair, the size of which is $15,354 \times 32,507$ pixels [33], we cropped the original image into small-sized images with a size of 256×256 pixels for deep learning. We divided the cropped images for use in the training set, validation set, and test set with the ratio of 8:1:1. The cropped dataset contains 6,096 images for training, 762 images for test, and 762 images for validation. Fig.5 shows example images of the cropped BCDD dataset, Fig.5(a) and (b) show bitemporal image pairs, and Fig.5(c) shows the ground truth images.

B. IMPLEMENTATION DETAILS

The encoder of the proposed network was designed based on VGG16 [37]. We used the first four convolutional blocks of VGG16 as the first four convolutional blocks of our encoder. Additionally, we used the weights of the pre-trained VGG16 with the ImageNet dataset [38] as the initial weights of the first four convolutional blocks of the encoder.

We set the number of filters K in Equation (3) to be the same as the number of channels C of the input for the channel-wise co-attention module, the number of filters L in Equation (5) to C, and the balance weight w and margin m for the contrastive loss function to 0.4 and 2.0, respectively. The balanced weight λ between the two losses in Equation (7) is set to 0.1 for the CDD and BCDD dataset.

We implemented DASNet and STANet using the PyTorch [39] codes provided by the authors without modifying the network structures [7], [22]. We implemented FC-EF, FC-Siam-Conc, FC-Siam-Diff, and the proposed method using the TensorFlow2 library [40]. We trained the proposed method using the Adam optimizer with a fixed learning rate of 1×10^{-4} on an NVIDIA TITAN XP graphics card. We set the batch size to 8, maximum epoch to 200, weight decay to 1×10^{-4}, and patience for early stopping to 30. In addition, we do not perform data augmentation for the CDD dataset or BCDD dataset.

C. PERFORMANCE METRICS

To evaluate the performance of the CD methods, we analyzed the precision, recall, f1-score, and overall accuracy. The precision is defined as

$$P = \frac{TP}{TP + FP}, \tag{8}$$

where TP is the number of true positives and FP is the number of false positives. The precision is the ratio of the number of pixels correctly classified as changed pixels to the number of pixels detected as changed pixels. We define recall as

$$R = \frac{TP}{TP + FN}, \tag{9}$$

where FN is the number of false negatives. The recall is the ratio of the number of pixels correctly classified as changed pixels to the total number of actually changed pixels. We define the f1-score as

$$F = \frac{2PR}{P + R}, \tag{10}$$

where F is the f1-score, P is precision, and R is recall. We define the overall accuracy as

$$OA = \frac{TP + TN}{TP + TN + FP + FN}, \tag{11}$$

where OA is overall accuracy and TN is the number of true negatives. We compute all the metrics in pixel units in this investigation.

D. EXPERIMENTAL RESULTS

1) CDD DATASET

We present the precision, recall, f1-score, and overall accuracy of each method for the CDD dataset in Table 1. As shown in the table, the proposed method achieves the best performance with the highest recall (95.67%), precision (96.06%), f1-score (95.86%), and overall accuracy (98.95%). The recall, precision, f1-score, and overall accuracy are approximately 2.23%, 5.91%, 4.10%, and 1.09%, respectively, higher than those of STANet, which achieves the second-best performance. The proposed method demonstrates a significant performance improvement compared with other methods in terms of both recall and precision. We believe that the reason for the improvement in the precision is that the proposed method may reduce the differences in unchanged areas by comparing a feature map in one image with similar feature maps in the other image. In addition, we think that the proposed method achieves the best performance without compromising the probability of detection because the channel-wise co-attention module also considers the difference between the feature maps from the two images, as shown in Equation (5).

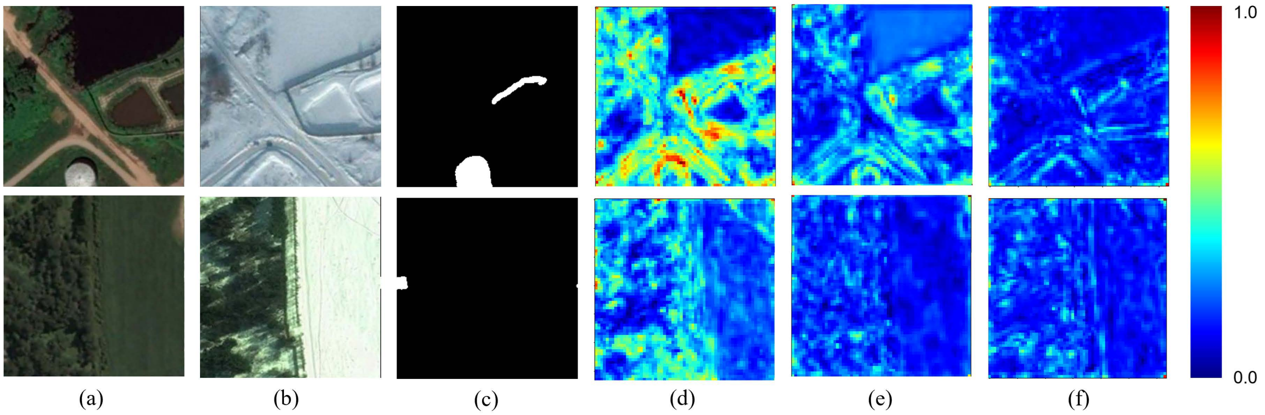

FIGURE 6. Illustrations of the results of the channel-wise co-attention module for the CDD dataset [6]: (a) T_0 image, (b) T_1 image, (c) ground truth, (d) mean(abs($F_0 - F_1$)), (e) mean(abs($F_0 - \tilde{F_1}$)), and (f) mean(abs($\tilde{F_0} - F_1$)).

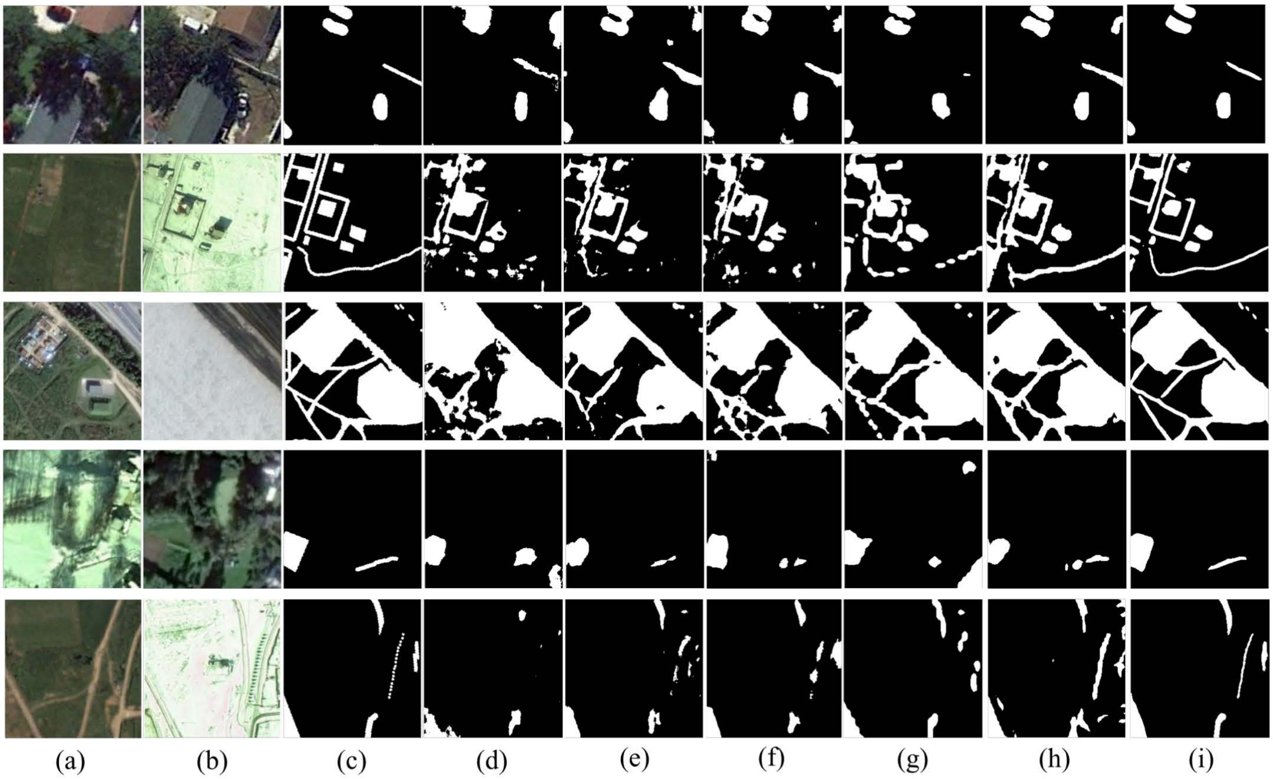

FIGURE 7. Comparison of the proposed method with other CD methods for the CDD dataset [6]: (a) T_0 image, (b) T_1 image, (c) ground truth, (d) FC-EF, (e) FC-Siam-Diff, (f) FC-Siam-Conc, (g) DASNet, (h) STANet, and (i) the proposed method. The changed parts are shown in white.

To further verify the effectiveness of the channel-wise co-attention module, we compared the absolute difference between the feature maps from the two images (*i.e.*, abs $(F_0 - F_1)$) with the absolute difference between the feature maps from an image and feature maps generated from the attention module (*i.e.*, abs$(F_0 - \tilde{F_1})$ and abs$(\tilde{F_0} - F_1)$). If the channel-wise co-attention module is effective in reducing differences between the feature maps from the unchanged areas, the absolute difference between F_0 and $\tilde{F_1}$ and between $\tilde{F_0}$ and F_1 will have smaller values than the absolute difference between F_0 and F_1. Fig.6 shows the absolute difference map averaged in the channel direction for visualization. As can be observed, for the unchanged areas, the absolute difference maps between F_0 and $\tilde{F_1}$ and between $\tilde{F_0}$ and F_1 have smaller values compared with the absolute difference map between F_0 and F_1. From these results, we can confirm that the channel-wise co-attention module is effective in alleviating the problem of pseudo-change detection.

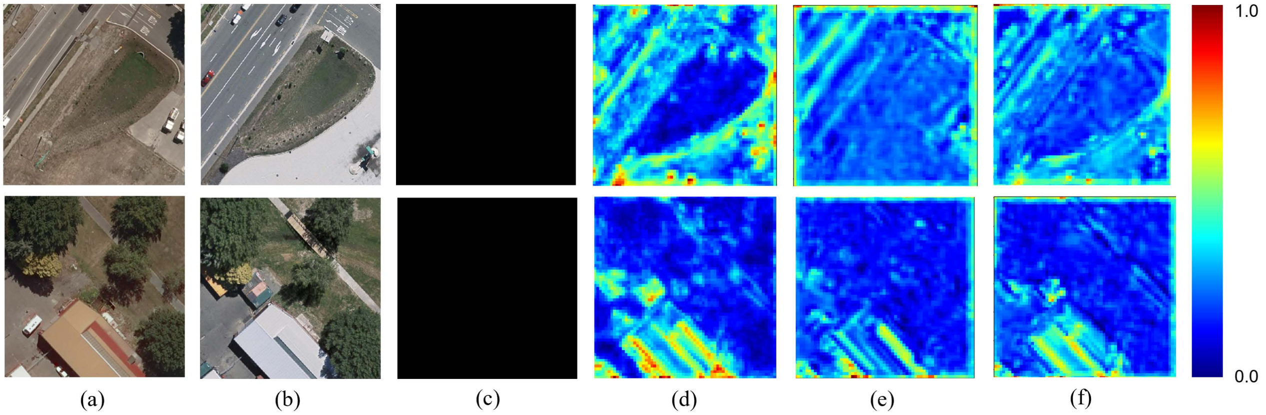

FIGURE 8. Illustrations of the results of the channel-wise co-attention module for the BCDD dataset [33]: (a) T_0 image, (b) T_1 image, (c) ground truth, (d) mean(abs($F_0 - F_1$)), (e) mean(abs($F_0 - \tilde{F_1}$)), and (f) mean(abs($\tilde{F_0} - F_1$)).

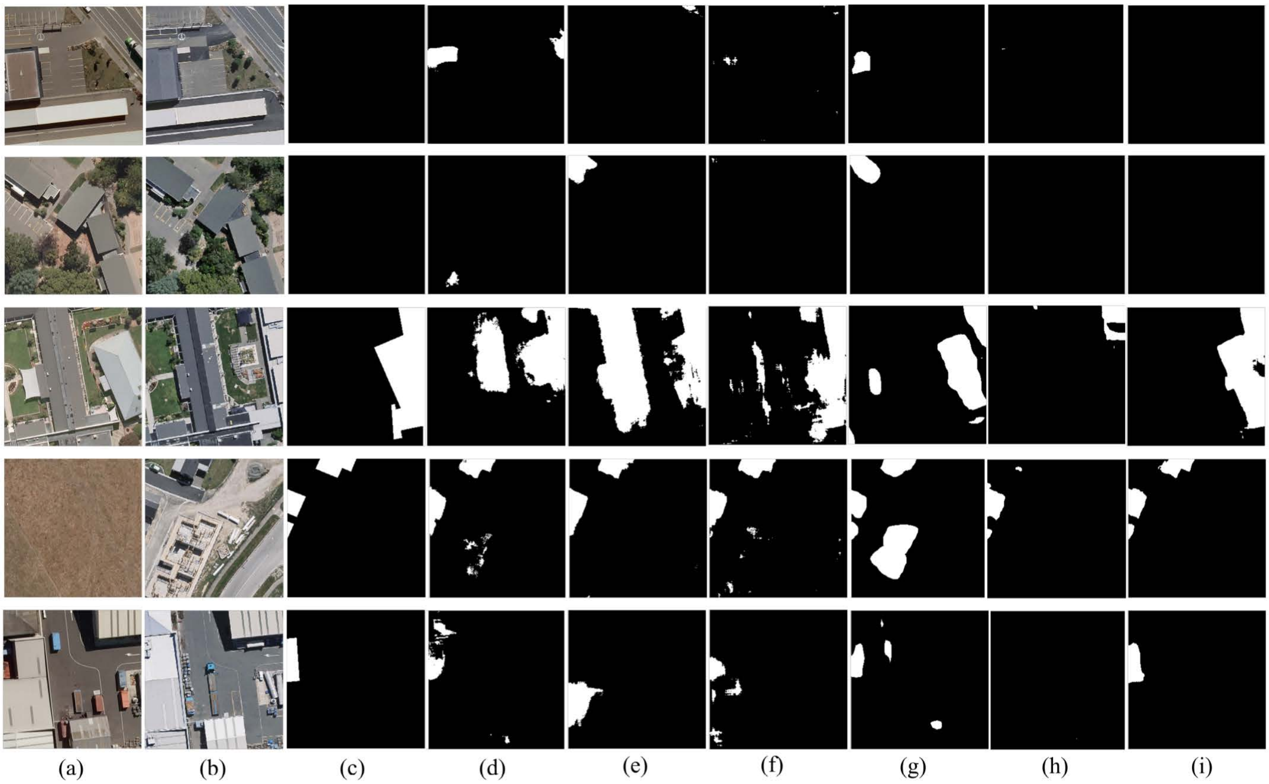

FIGURE 9. Comparison of the proposed method with other CD methods for the BCDD dataset [33]: (a) T_0 image, (b) T_1 image, (c) ground truth, (d) FC-EF, (e) FC-Siam-Diff, (f) FC-Siam-Conc, (g) DASNet, (h) STANet, and (i) the proposed method. The changed parts are shown in white.

In addition, we think that the contrastive loss function is effective in reducing pseudo-change detection because it forces the features from the unchanged areas of the two images to be similar. The contrastive loss function was also used in previous methods such as STANet and DASNet [7], [22], which may explain why STANet and DASNet performed better than the other conventional methods.

For a more intuitive evaluation, Fig.7 compares the detection results of the proposed method with other CD methods. Figs.7(a) and (b) show bitemporal images affected by illumination changes and seasonal changes, Fig.7(c) shows the ground truth images, and Figs.7(d), (e), (f), (g), (h), and (i) show the detection results of FC-EF, FC-Siam-Diff, FC-Siam-Conc, DASNet, STANet, and the proposed method, respectively. As can be observed, the result of the proposed method is the most similar to the ground truth, which means that the proposed method is more robust to environmental pseudo-changes than other methods. In addition,

TABLE 1. Performance results for the CDD dataset.

Method	R (%)	P (%)	F (%)	OA (%)
FC-EF	89.54	87.12	88.32	96.98
FC-Siam-Diff	93.16	87.92	90.46	97.49
FC-Siam-Conc	93.34	87.93	90.55	97.51
DASNet	93.41	90.13	91.74	97.86
STANet	93.44	90.15	91.76	97.86
proposed method	**95.67**	**96.06**	**95.86**	**98.95**

we qualitatively confirm that a channel-wise co-attention module may be more effective in reducing pseudo-changes than self-attention and spatial-wise co-attention modules, as shown in Figs.7(g), (h), and (i).

2) BCDD DATASET

To further evaluate the performance of the proposed method, we also conducted experiments using the BCDD dataset. We present the precision, recall, f1-score, and overall accuracy in Table 2.

TABLE 2. Performance results for the BCDD dataset.

Method	R (%)	P (%)	F (%)	OA (%)
FC-EF	84.99	87.68	86.31	98.70
FC-Siam-Diff	87.32	65.73	75.00	97.18
FC-Siam-Conc	85.88	60.98	71.32	96.66
DASNet	89.34	88.88	89.11	98.94
STANet	89.26	74.73	81.35	98.34
proposed method	**91.11**	**95.03**	**93.03**	**99.34**

As shown in Table 2, the proposed method also achieves the best performance with the highest recall (91.11%), precision (95.03%), f1-score (93.03%), and overall accuracy (99.34%). The recall, precision, f1-score, and overall accuracy are approximately 1.77%, 6.15%, 3.92%, and 0.4%, respectively, higher than those of DASNet, which achieves the second-best performance. In particular, the proposed method demonstrates significant performance improvement compared with other methods in terms of precision, which implies the proposed method is effective in reducing the detection of pseudo-changes.

We also identify whether the proposed method can reduce differences between feature maps when comparing feature maps from unchanged regions of two images with environmental changes. Fig.8 compares the absolute difference between the feature maps from two images (*i.e.,* abs($F_0 - F_1$)) with the absolute difference between feature maps from an image and feature maps generated from the attention module (*i.e.,* abs($F_0 - \tilde{F_1}$) and abs($\tilde{F_0} - F_1$)). As shown in Figs.8(d), (e), and (f), for the unchanged areas, the absolute differences between F_0 and $\tilde{F_1}$ and between $\tilde{F_0}$ and F_1 have smaller values than the absolute difference between F_0 and F_1. From this figure, we can observe that the proposed method reduces the difference between feature maps from unchanged areas, which is consistent with the experiments using the CDD dataset.

Fig.9 illustrates the prediction results of each method using the BCDD dataset. Figs.9(a) and (b) show bitemporal images, Fig.9(c) shows the ground truth images, and Figs.9(d), (e), (f), (g), (h), and (i) show the detection results of FC-EF, FC-Siam-Diff, FC-Siam-Conc, DASNet, STANet, and the proposed method, respectively. From this figure, we also confirm that the proposed method is effective in reducing the detection of pseudo-changes.

V. CONCLUSION

In this study, we propose a channel-wise co-attention-based Siamese network system to detect changes between high-resolution bitemporal images. Compared with existing change detection methods, the proposed method is more robust to pseudo-changes caused by different imaging conditions and/or changes irrelevant to the purposes of change detection. The key element of the proposed method for reducing pseudo-change detection is the channel-wise co-attention module that considers channel-wise correlations between one feature map in an image and feature maps from the other image to find similar feature maps in the other image. By comparing the feature map in one image with the combination of similar feature maps in the other image instead of comparing the same channel feature maps, the proposed method reduces the detection of pseudo-changes. In addition, the contrastive loss function of the proposed method encourages the features of the two images from unchanged areas to be more similar, thereby facilitating the determination of unchanged areas as unchanged areas, also alleviating the problem of pseudo-change detection. We verified that the proposed method is more robust to pseudo-changes than conventional methods through experiments using the change detection dataset (CDD) [6] and building change detection dataset (BCDD) [33]. The proposed method achieves significant performance improvement compared with existing methods in terms of both recall and precision as demonstrated in the experiments.

REFERENCES

[1] L. Khelifi and M. Mignotte, "Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis," *IEEE Access*, vol. 8, pp. 126385–126400, 2020.

[2] X. Huang, L. Zhang, and T. Zhu, "Building change detection from multitemporal high-resolution remotely sensed images based on a morphological building index," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 1, pp. 105–115, Jan. 2014.

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

[4] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, "Change detection in synthetic aperture radar images based on deep neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 1, pp. 125–138, 2017.

[5] S. Mahdavi, B. Salehi, W. Huang, M. Amani, and B. Brisco, "A PolSAR change detection index based on neighborhood information for flood mapping," *Remote Sens.*, vol. 11, no. 16, p. 1854, Aug. 2019.

[6] M. Lebedev, Y. V. Vizilter, O. Vygolov, V. Knyaz, and A. Y. Rubis, "Change detection in remote sensing images using conditional adversarial networks," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 42, no. 2, pp. 565–571, 2018.

[7] J. Chen, Z. Yuan, J. Peng, L. Chen, H. Huang, J. Zhu, Y. Liu, and A. H. Li, ''DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images,'' *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1194–1206, 2021.

[8] P. de Bem, O. de Carvalho Junior, R. F. Guimarães, and R. T. Gomes, ''Change detection of deforestation in the Brazilian Amazon using Landsat data and convolutional neural networks,'' *Remote Sens.*, vol. 12, no. 6, p. 901, Mar. 2020.

[9] A. Lefebvre, T. Corpetti, and L. Hubert-Moy, ''Object-oriented approach and texture analysis for change detection in very high resolution images,'' in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, vol. 4, Jul. 2008, p. IV-663.

[10] M. Bouziani, K. Goïta, and D.-C. He, ''Automatic change detection of buildings in urban environment from very high spatial resolution images using existing geodatabase and prior knowledge,'' *ISPRS J. Photogramm. Remote Sens.*, vol. 65, no. 1, pp. 143–153, Jan. 2010.

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

[12] C. Zhang, G. Li, and W. Cui, ''High-resolution remote sensing image change detection by statistical-object-based method,'' *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 7, pp. 2440–2447, Jul. 2018.

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

[14] P. Du, X. Wang, D. Chen, S. Liu, C. Lin, and Y. Meng, ''An improved change detection approach using tri-temporal logic-verified change vector analysis,'' *ISPRS J. Photogramm. Remote Sens.*, vol. 161, pp. 278–293, Mar. 2020.

[15] M. H. Kesikoglu, Ü. H. Atasever, and C. Özkan, ''Unsupervised change detection in satellite images using fuzzy c-means clustering and principal component analysis,'' *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 7, pp. 129–132, Oct. 2013.

[16] 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, Aug. 2009.

[17] V. Badrinarayanan, A. Kendall, and R. Cipolla, ''SegNet: A deep convolutional encoder-decoder architecture for image segmentation,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[18] O. Ronneberger, P. Fischer, and T. Brox, ''U-Net: Convolutional networks for biomedical image segmentation,'' in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[19] G. Koch, ''Siamese neural networks for one-shot image recognition,'' in *Proc. ICML Deep Learn. Workshop*, Lille, France, vol. 2, 2015, pp. 1–30.

[20] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, and F. Porikli, ''See more, know more: Unsupervised video object segmentation with co-attention Siamese networks,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3623–3632.

[21] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, ''Dual attention network for scene segmentation,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[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, May 2020.

[23] 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, pp. 1–16, 2022.

[24] 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.

[25] Y. Zhang, S. Zhang, Y. Li, and Y. Zhang, ''Coarse-to-fine satellite images change detection framework via boundary-aware attentive network,'' *Sensors*, vol. 20, no. 23, p. 6735, Nov. 2020.

[26] A. Varghese, J. Gubbi, A. Ramaswamy, and P. Balamuralidhar, ''ChangeNet: A deep learning architecture for visual change detection,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018, pp. 1–16.

[27] H. Jiang, X. Hu, K. Li, J. Zhang, J. Gong, and M. Zhang, ''PGA-SiamNet: Pyramid feature-based attention-guided Siamese network for remote sensing orthoimagery building change detection,'' *Remote Sens.*, vol. 12, no. 3, p. 484, Feb. 2020.

[28] 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 J. Photogramm. Remote Sens.*, vol. 166, pp. 183–200, Aug. 2020.

[29] Y. Lei, X. Liu, J. Shi, C. Lei, and J. Wang, ''Multiscale superpixel segmentation with deep features for change detection,'' *IEEE Access*, vol. 7, pp. 36600–36616, 2019.

[30] H. Chen, C. Wu, and B. Du, ''Towards deep and efficient: A deep Siamese self-attention fully efficient convolutional network for change detection in VHR images,'' 2021, *arXiv:2108.08157*.

[31] Q. Guo, J. Zhang, S. Zhu, C. Zhong, and Y. Zhang, ''Deep multiscale Siamese network with parallel convolutional structure and self-attention for change detection,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.

[32] H. Hao, S. Baireddy, E. R. Bartusiak, L. Konz, K. LaTourette, M. Gribbons, M. Chan, M. L. Comer, and E. J. Delp, ''An attention-based system for damage assessment using satellite imagery,'' 2020, *arXiv:2004.06643*.

[33] 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.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, ''Attention is all you need,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998-6008.

[35] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, ''Self-attention generative adversarial networks,'' in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7354–7363.

[36] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, ''Encoder-decoder with atrous separable convolution for semantic image segmentation,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[37] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' 2014, *arXiv:1409.1556*.

[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, ''ImageNet: A large-scale hierarchical image database,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[39] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, and L. Antiga, ''Pytorch: An imperative style, high-performance deep learning library,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 8026–8037.

[40] M. Abadi *et al.*, ''TensorFlow: Large-scale machine learning on heterogeneous distributed systems,'' 2016, *arXiv:1603.04467*.

EUNJEONG CHOI received the B.S. and M.S. degrees in electronic engineering from Ewha Womans University, Seoul, South Korea, in 2016 and 2018, respectively, where she is currently pursuing the Ph.D. degree in electronic and electrical engineering. Her research interests include deep learning for machine vision and digital signal processing.

JEONGTAE KIM received the B.S. and M.S. degrees in control and instrumentation engineering from Seoul National University, Seoul, South Korea, in 1989 and 1991, respectively, and the Ph.D. degree in electrical engineering and computer science from the University of Michigan, Ann Arbor, in 2004. From 1991 to 1998, he had worked at Samsung Electronics, South Korea, where he had been engaged in the development of digital camcorder and digital TV. Since 2004, he has been with the Department of Electronic and Electrical Engineering, Ewha Womans University, Seoul, as a Professor. His research interests include machine learning, computer vision, and radar signal processing.