Local Similarity Siamese Network for Urban Land Change Detection on Remote Sensing Images

Haeyun Lee, *Student Member, IEEE*, Kyungsu Lee, *Student Member, IEEE*, Jun Hee Kim, *Member, IEEE*, Younghwan Na*, Student Member, IEEE*, Juhum Park*, Member, IEEE*, Jihwan P. Choi, *Senior Member, IEEE*, and Jae Youn Hwang, *Member, IEEE*

***Abstract*—Change detection is an important task in the field of remote sensing. Various change detection methods based on convolutional neural networks (CNNs) have recently been proposed for remote sensing using satellite or aerial images. However, existing methods allow only the partial use of content information in images during change detection because they adopt simple feature similarity measurements or pixel-level loss functions to construct their network architectures. Therefore, when these methods are applied to complex urban areas, their performance in terms of change detection tends to be limited. In this article, a novel CNN-based change detection approach, referred to as a local similarity Siamese network (LSS-Net), with a cosine similarity measurement, was proposed for better urban land change detection in remote sensing images. To use content information on two sequential images, a new change attention map-based content loss function was developed in this study. In addition, to enhance the performance of the LSS-Net in terms of change detection, a suitable feature similarity measurement method, incorporated into a local similarity attention module, was determined through systemic experiments. To verify the change detection performance of the LSS-Net, it was compared with other state-of-the-art methods. The experimental results show that the proposed method outperforms the state-of-the-art methods in terms of the F1 score (0.9630, 0.9377, and 0.7751) and kappa (0.9581, 0.9351, and 0.7646) on the three test datasets, thus suggesting its potential for various remote sensing applications.**

***Index Terms*—Change detection, remote sensing, Siamese network, similarity attention.**

I. Introduction

REMOTE sensing has been widely used for high-throughput monitoring of areas that cannot be accessed nonintrusively [1]. Various studies have been conducted on the application of remote sensing techniques, such as remote sensing image classification and segmentation [2]–[6], to automatically obtain useful information of interest from remote sensing images. In particular, change detection is one of the most important tasks in remote sensing imagery analysis [7]. Thus far, change detection has been applied to a wide range of applications, such as monitoring of changes in vegetation, urban expansion, agriculture, disasters, illegal woodcutting, and the melting of polar icecaps [8]–[11].

The main objective of change detection for remote sensing is to detect significant changes from sequential remote sensing images acquired at different points in time. In change detection, variations in environmental conditions can result in challenges in extracting important information from remote sensing images. Therefore, several methods have been developed to address the challenging issues in change detection [12]–[14]. Benedek *et al.* [14] proposed a multilayer conditional mixed Markov model for change detection in aerial images. Chen *et al.* [12] used a multithreshold strategy for urban building change detection using both aerial and Lidar images. In addition, Bourdis *et al.* [13] devised an optical-flow-based change detection method for solving parallax problems. However, the aforementioned methods are heavily influenced by parameter tuning and the presence of other environmental factors, such as shadows and occluded objects.

Deep learning techniques have recently been applied to remote sensing images [15]–[19]. Liu *et al.* [17] proposed an unsupervised symmetric convolution coupling network (SCCN) using optical and radar images for change detection. In addition, Zhan *et al.* [16] developed a deep Siamese convolutional network with Euclidean distance for change detection. The features extracted from the deep Siamese convolutional network are more robust and abstract than hand-crafted features. It therefore exhibits excellent results compared to those based on classical approaches. However, it requires manual identification of the threshold for each image pair. Daudt *et al.* [15] examined three deep learning architectures: Fully convolutional early fusion (FC-EF), fully convolutional Siamese-concatenation (FC-Siam-conc), and fully convolutional Siamese-difference

Manuscript received December 3, 2020; revised February 11, 2021; accepted March 12, 2021. Date of publication March 26, 2021; date of current version April 26, 2021. This work was supported in part by the National Research Fundation of Korea (NRF) under Grant NRF-2020R1A2B5B01002786. The work of Jae Youn Hwang was supported in part by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) through the Korean Government (MSIT) under Grant NRF-2017M3A9G8084463. *(Corresponding authors: Jae Youn Hwang and Jihwan P. Choi.)*

Haeyun Lee, Kyungsu Lee, Younghwan Na, and Jae Youn Hwang are with the Information and Communication Engineering, Daegu Gyeongbuk Institute of Science and Technology, Daegu 42988, South Korea (e-mail: haeyun@dgist.ac.kr; ks_lee@dgist.ac.kr; nyh0426@dgist.ac.kr; jyhwang@dgist.ac.kr).

Jun Hee Kim is with the Agency for Defense Development, Daejoen 34186, South Korea (e-mail: kjh1127@add.re.kr).

Juhum Park is with the Dabeeo Inc., Seoul 04107, South Korea (e-mail: juhum.park@dabeeo.com).

Jihwan P. Choi is with the Department of Aerospace Engineering, Korea Advanced Institute of Science and Technology, Daejoen 34141, South Korea (e-mail: jhch@kaist.ac.kr).

Digital Object Identifier 10.1109/JSTARS.2021.3069242

(FC-Siam-diff). Moreover, Jiang *et al.* [19] proposed a building change detection network called PGA-SiamNet, which uses pyramid feature-based attention modules. These deep learning-based methods are less affected by environmental factors and thus show higher performance than classical approaches. Because these learning-based methods only utilize a binary cross-entropy loss function, which is a pixel-wise loss function, for training the models, they exhibited limitations in the use of content information, which may be key characteristics of remote sensing data for better change detection. In addition, they merely used a difference or concatenation to measure the similarity between the input images. Therefore, they are limited to comparing multiple levels of features for change detection in remote sensing images.

In this study, we propose a local similarity Siamese network (LSS-Net) to resolve the aforementioned problems for more efficient change detection in remote sensing images. First, our network uses a Siamese network architecture consisting of two identical subnetworks with the same parameters and weights. We then adopt a local similarity module, capable of using each low- and high-level feature, to detect changes in the two input images. Because the contents of an image can be divided into low- and high-level features [20], we configured the network to use the features at each level accordingly. In addition, we propose a new change attention map-based content loss function to efficiently use content information on images during the training of the LSS-Net and therefore achieve better optimization. Note that the loss functions using content information have thus far been applied only to image synthesis and generation tasks. However, in this study, we demonstrated that content information-based loss is effective for high-level vision tasks.

Our main contributions are summarized as follows.

1) We constructed a novel deep learning model, denoted as LSS-Net, which outperforms other state-of-the-art methods on several remote sensing datasets for change detection.
2) We also proposed a change attention map-based content loss (*CAC Loss*) for using content information from two sequential images. We presented an ablation study and qualitative results to demonstrate the effectiveness of the *CAC Loss*. The *CAC Loss* improved the performance of all change detection networks, including our method, from at least 0.002 to a maximum of 0.0394 in terms of the F1 score by using the content information of the input images.
3) We investigated the effect of a local similarity attention module and a decoder structure on change detection, thus showing that these utilities improve the performance of the deep learning network for such detection in two ways: a) Through an ablation study, we determined a similarity measurement method suitable for change detection in urban areas and applied the method to the attention module. b) The local similarity attention module for low-level features detects the edges of the changed objects, and the LSS-Net then produces better shapes based on the change-detection results.

II. Methods

In this section, we first introduce our network architecture for change detection, which we refer to as LSS-Net. Subsequently, we introduce a local similarity attention module. Finally, we describe the training procedure for LSS-Net and the change attention map-based content loss function proposed herein.

A. Network Architecture

Fig. 1 shows the architecture of our network, LSS-Net. The network takes two remote-sensing images of the same location acquired at different times as inputs and then outputs a change-prediction map. Our network, LSS-Net, consists of three parts: An encoder for feature extraction, local similarity attention modules for calculating the similarity of the features, and a decoder. For the encoder part, we adopted SE-ResNet-50, which is a squeeze-and-excitation residual neural network [21]. Because the squeeze-and-excitation module can emphasize informative channels [21], this module can highlight the channels required for change detection. To exploit various high- and low-level features for change detection, we divided the level of convolution blocks into five different levels, in which each level represented a group of residual units sharing the same spatial resolution, as shown in Fig. 1. Thus, the network can use five feature maps with different spatial resolutions. We propose a local similarity attention module to calculate the similarity of features. We then obtain the similarity by dividing the high- and low-level features [20], [22]. This module is described in detail in Section II-B. The decoder combines the results obtained by calculating the similarity of the features at each level and then produces the final output. In the decoder, we use one concatenation, two convolution layers, and two pixel-shuffles [23] to upsample the feature maps, as shown in Fig. 1. Before passing through the decoder, a high-level feature map conducts a four times upsampling to match the spatial resolution of the low-level feature map.

B. Local Similarity Attention Module

In this subsection, we describe the local similarity attention module used in the LSS-Net. Previous change detection methods using an attention module merely use channel-wise information and an insufficient amount of spatial information on the input images [19], [24]. In contrast, we devised a local similarity attention module to fully utilize the spatial information on input images. As shown in Fig. 1, in each similarity attention module, input features were divided into high- and low-level features; the features from each image were then concatenated and input to the next channel. For high-level features, level 3, 4, and 5 feature maps were used. To concatenate these feature maps, the level 4 and 5 feature maps were upsampled up to two and four times, respectively. Likewise, level 1 and 2 feature maps were used for the low level. The level 2 feature map was upsampled twice to concatenate with the level 1 feature map. To calculate the similarity, we used the attention module-based cosine similarity. The similarity attention value, SA, was obtained from the

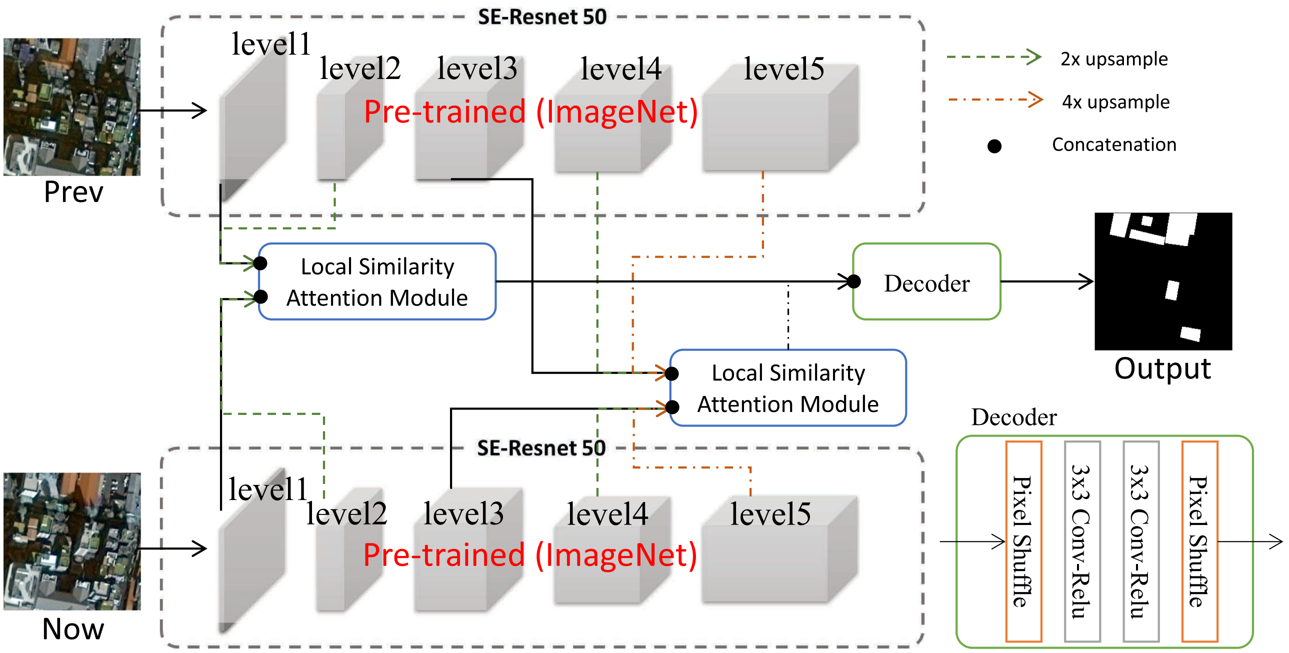


Fig. 1. Proposed network architecture, LSS-Net, for change detection. This network uses the Siamese network architecture to obtain features from two remote sensing images at different times. The proposed local similarity attention modules were used to calculate the similarity of high- and low-level features. Finally, the decoder produces the result of change detection by integrating high- and low-level features.

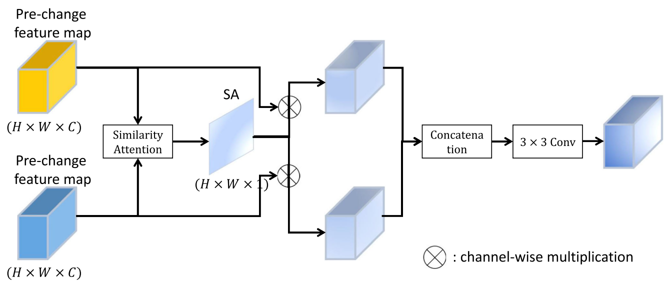


Fig. 2. Local similarity attention module for change detection. This attention module uses cosine similarity to extract spatial information from the input images.

following:

$$SA(i,j) = 1 - \frac{\sum_{k=1}^{C} F_{\text{prev}}^{k}(i,j) \times F_{\text{now}}^{k}(i,j)}{\sqrt{\sum_{k=1}^{C} \left(F_{\text{prev}}^{k}(i,j)\right)^2} \times \sqrt{\sum_{k=1}^{C} \left(F_{\text{now}}^{k}(i,j)\right)^2}} \quad (1)$$

where C is the number of channels of the feature maps, F_{prev}^{k} and F_{now}^{k} are the feature maps of channel k of the pre- and post-change input images, respectively, and i and j are the spatial coordinates of the feature maps and similarity attention values, respectively. The SA obtained through this procedure was multiplied by each input feature map. The feature maps multiplied by SA were then concatenated and passed through a 3×3 convolution filter to apply the local information and adjust the number of channels. Fig. 2 visualizes the local similarity attention module described.

C. Training

In this subsection, we describe how we trained the LSS-Net with the proposed change attention map-based content loss function.

a) Loss function: To train the LSS-Net, we used two loss functions: A binary cross-entropy loss function and a change attention map-based content loss. However, conventional semantic segmentation loss functions, such as binary cross-entropy loss, are considered for each pixel only when training a network. Therefore, the surrounding information may not be properly used because each pixel is compared. In particular, incorrect results may be derived from remote sensing images when the images are not orthorectified. This loss function is not therefore suitable for reflecting contextual information.

Thus, we propose the *CAC loss* to solve this problem. The *CAC loss* developed herein is inspired by another content loss function from a previous study [25]. To calculate the *CAC loss*, CA_{map} is first obtained as follows:

$$CA_{\text{map}} = \text{sigmoid}\left(\hat{J}_0 - \hat{J}_1\right) \quad (2)$$

where $\hat{J}_c$ is the logit of the network at the cth channel. $\hat{J}_0$ and $\hat{J}_1$ represent the logit for the unchanged and changed labels, respectively. We use the sigmoid activation function to normalize the value of CA_{map} from zero to one. Here, sigmoid$(\hat{J}_0 - \hat{J}_1)$ signifies that the unchanged part has a value close to one, whereas the changed part has a value close to zero.

Given a training dataset $D = \{\ldots, (I_{\text{prev}}^{(i)}, I_{\text{now}}^{(i)}, GT^{(i)}), \ldots\}$, where $I_{\text{prev}}^{(i)}$ and $I_{\text{now}}^{(i)}$ are the ith remote sensing images acquired at the prechange and postchange time points, respectively, and

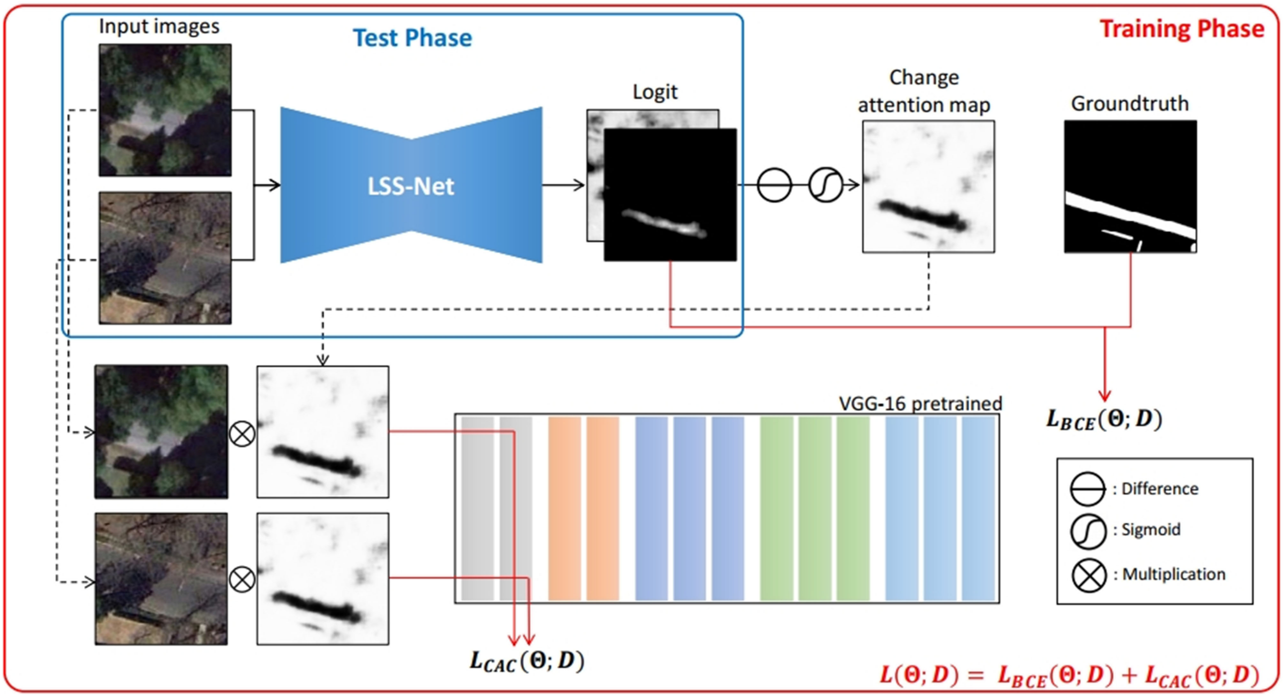


Fig. 3. Training and test procedures. The combination of the binary cross-entropy loss, $L_{\text{BCE}}(\Theta; D)$, and change attention map-based loss functions, $L_{\text{CAC}}(\Theta; D)$, was used to train the LSS-Net. In contrast, in the test phase, only the trained LSS-Net, indicated by a blue solid box, was used to predict changes in the images.

$GT^{(i)}$ is the corresponding ground truth. The *CAC loss* can be calculated as follows:

$$\begin{aligned} L_{CAC}(\Theta; D) = \sum_i \| & VGG_{\text{relu1\_2}}(CA_{\text{map}} * I^{(i)}_{\text{prev}}) \\ & - VGG_{\text{relu1\_2}}(CA_{\text{map}} * I^{(i)}_{\text{now}}) \|^2 \end{aligned} \tag{3}$$

where Θ represents the parameters of the network, $*$ is a pixel-wise multiplication for each channel, and $\text{VGG}_{\text{relu1\_2}}$ is a loss network for the use of the content information of two sequential images. We adopted the loss network, $\text{VGG}_{\text{relu1\_2}}$, of VGG16 [26] pretrained on ImageNet data [27] and compute the *CAC loss* in the relu*1_2* layer. Fig. 3 shows our training procedure and the *CAC loss* function. The following loss function (4) is used to train the LSS-Net

$$L(\Theta; D) = \alpha \times L_{\text{BCE}}(\Theta; D) + L_{\text{CAC}}(\Theta; D) \tag{4}$$

where $L_{\text{BCE}}(\Theta; D)$ is a pixel-wise segmentation loss function (binary cross-entropy loss), and α is a weighting factor. The value of α was determined to be one through. In Section III-A, we describe additional experiments for our *CAC loss* function.

b) Dataset: For the training and testing of the LSS-Net, we used two public datasets: ChangeDetectionDataset (CDD) obtained from Google Earth (DigitalGlobe) [28], and the Wuhan University (WHU) building change detection dataset [29]. We also created a satellite image change detection dataset, denoted as an urban change-detection (UCD) dataset. Fig. 4 shows an example of each dataset.

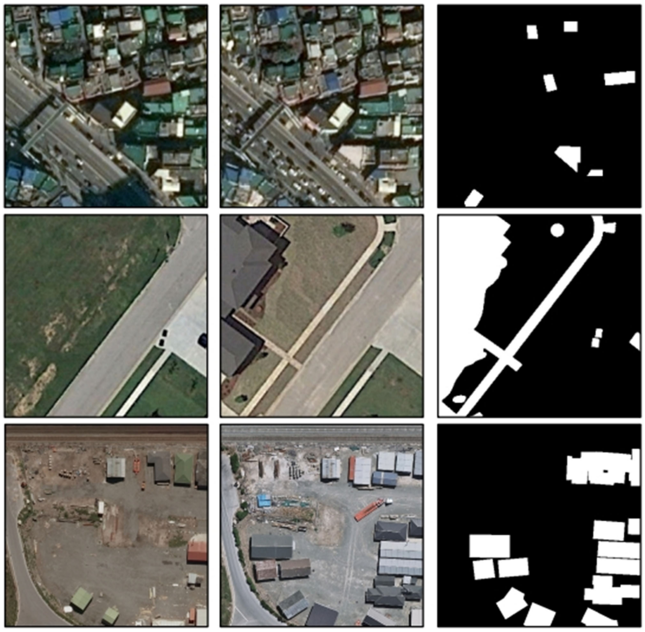


Fig. 4. Examples from change detection dataset: Middle left, input images; right, ground truth of the change detection (first row, UCD dataset; second row, CDD dataset [28]; and third row, Wuhan University (WHU) building change detection dataset [29]).

The CDD dataset was cropped into 256 × 256 pixels with several augmentations [28]. Its spatial resolution ranges from 3 to 100 cm/px. The dataset comprised 10 000 training image pairs, 3000 validation image pairs, and 3000 test image pairs. The images on the CDD dataset showed a sparse distribution of objects, such as buildings and roads, as can be seen in the second row in Fig. 4.

The WHU dataset covers Christchurch, New Zealand, and comprises two images captured in 2012 and 2016, as well as the detection labels of the changed buildings. The image of the WHU dataset is an aerial image with a spatial resolution of 7.5 cm/px. The image size of the WHU dataset is 32507 × 15354 pixels. We cropped these image pairs into 256 × 256 pixels without overlapping. We randomly divided the dataset into training and test image pairs. In addition, we applied data augmentation such as random horizontal/vertical flips and random rotation by 90°. The number of training and testing image pairs were 6525 and 725, respectively. The CDD and WHU datasets are the most commonly used datasets for change detection on remote sensing images.

The UCD dataset was cropped into a pixel resolution of 256 × 256. We applied the same augmentation to the WHU dataset. The spatial resolution of an image of the UCD dataset was 50 cm/px. This dataset comprised 4400 training image pairs and 155 test image pairs. The images in the UCD dataset showed extremely complex urban areas, including numerous buildings and complicated roads.

c) Training setting: For the training of the LSS-Net, we initialized an encoder using SE-ResNet50, pretrained by applying ImageNet data, and a decoder using a Kaiming initializer [30]. We used PyTorch [31] to implement and train all the models, including the LSS-Net. In addition, we used the Adam optimizer [32], with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$, set the learning rate to 0.0001, and reduced it by one-half every 30 epochs. We used a minibatch size of 16 and trained the models for 60 epochs. Finally, we trained all the models on an Intel Xeon E5-2620 v4 @ 2.10 GHz and an NVIDIA TITAN RTX (24 GB).

III. Experiments

A. Ablation Study of Similarity Measurement and Loss Function

In this subsection, we discuss the investigation of the similarity measurement and proposed a loss function. In all the experiments, the performances of different network architectures with different similarity measurement methods and loss functions were evaluated using the UCD dataset. We performed experiments using four similarity measurement methods, namely, the difference, concatenation, correlation, and cosine similarity, to determine the final network architecture. The difference and concatenation were applied to subtract and concatenate the feature maps of two sequential images. We also used the correlation operation proposed in [34] and the cosine similarity mentioned in Section II-B. Table I outlines the experimental results of the similarity measurements. Among the different similarity measurements, the cosine similarity yielded the highest F1 score to be achieved. Therefore, we used the cosine similarity in a local similarity attention module of our network. In addition, a qualitative result of the similarity measurements is shown in Fig. 5. As indicated in Fig. 5, the cosine similarity yielded better results than the other similarity measurements.

TABLE I
Comparison of Similarity Measurement Methods on the UCD Dataset

Similarity Measurement Method	Difference	Concat	Correlation	Cosine
F1 score	0.7289	0.7309	0.6931	**0.7505**

Best performance is indicated in **bold.**

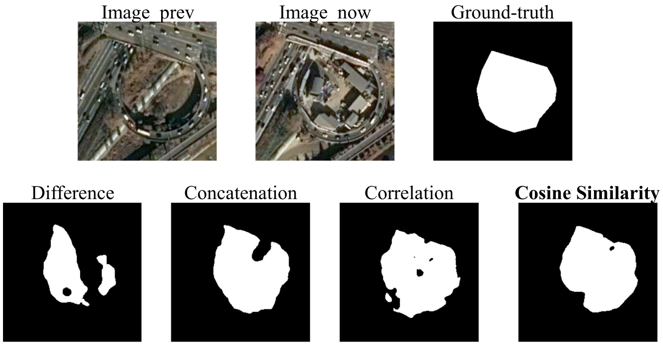


Fig. 5. Qualitative comparison of similarity measurement on the UCD dataset. The cosine similarity performs better than the other similarity measurements.

TABLE II
Comparison of Performance of the LSS-Net With Different Implementation Layers of *CAC Loss* on the UCD Dataset

Feature	None	relu1_2	relu2_2	relu3_3	relu4_3	relu5_3
F1 score	0.7505	**0.7751**	0.7720	0.7550	0.7358	0.7591

Best performance is indicated in **bold.**

We then conducted additional experiments to assess *CAC loss*. During the experiments, cosine similarity was used as the similarity measurement method. We first examined five different features, namely, *relu1_2*, *relu2_2*, *relu3_3*, *relu4_3*, and *relu5_3*, to determine the implementation layer of *CAC loss* in the pretrained VGG16 and then compared the outcomes obtained when *CAC loss* was used and when it was not (with only a binary cross entropy loss). Table II shows the results of the experiments on the *CAC loss*. In most cases shown in Table II, except for *relu4_3*, *CAC loss* offers better performance than using binary cross-entropy (BCE) loss alone. In particular, when the *relu1_2* feature of the loss network was used, the F1 score was improved by 0.7751, which is 0.0246 higher than that without *CAC loss*. Therefore, we trained our final network using the *relu1_2* feature. We also added the qualitative results of this ablation study in Fig. 6.

In addition, we investigated whether the *CAC loss* can improve the performance of other methods in change detection. We trained Siamese structure-based deep learning methods using *CAC loss*. Table III demonstrates that the performance of all the methods improved with the use of the *CAC loss*. In particular, the performance of DSMS-FCN increased by 0.0394 in terms of the F1 score. Through this experiment, we demonstrated that the *CAC loss* is effective for all the methods.

TABLE III
EFFECT OF OUR CHANGE ATTENTION-BASED CONTENT (CAC) LOSS FUNCTION ON OTHER MODELS

Method (F1 scroe)	FC-Siam-conc [15]	FC-Siam-diff [15]	DSMS-FCN [33]	FCN-PP [18]	PGA-SiamNet [19]	LSS-Net
BCE loss	0.6722	0.6246	0.6295	0.7035	0.7227	0.7505
BCE loss + CAC loss	0.6897	0.6591	0.6689	0.7055	0.7358	0.7751
Margin	+ 0.0175	+ 0.0345	+ 0.0394	+ 0.0020	+ 0.0174	+ 0.0246

We compare the performance of the LSS-Net and other models in Terms of the **F1 Score** using only binary cross-entropy (BCE) loss and change attention-based content (CAC) loss.

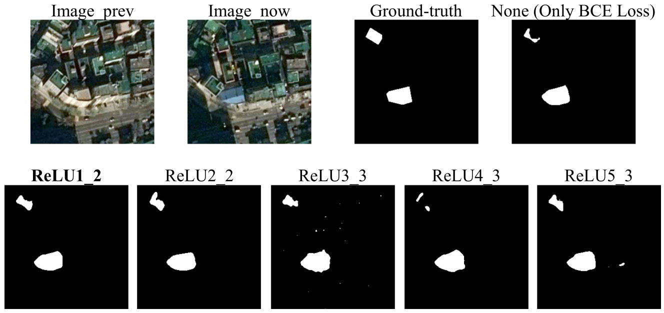


Fig. 6. Qualitative comparison of a CAC loss function on UCD dataset. The relu1_2 feature for CAC loss performs better than the other features as well as BCE loss.

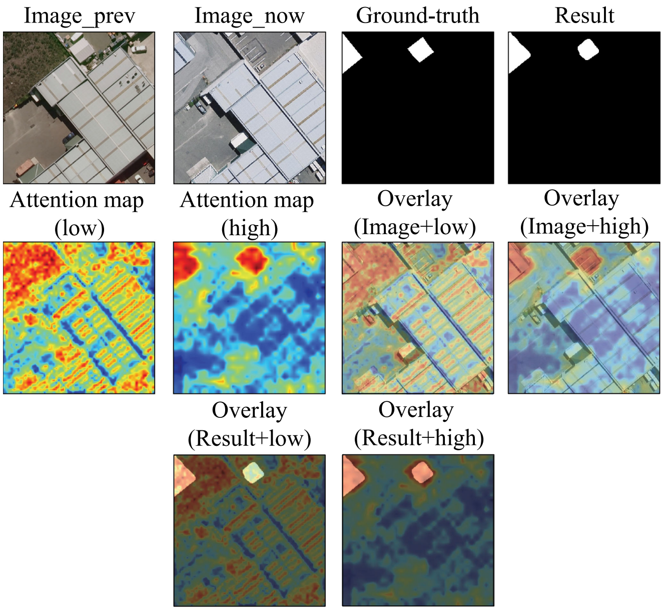


Fig. 7. Visualization of attention maps of a local similarity attention module on WHU dataset [29].

B. Analysis of Local Similarity Attention Module

In this subsection, to present our investigation of how high- and low-level local similarity attention modules affect change detection, we analyze each local similarity attention module. Fig. 7 shows the attention maps of the local similarity attention module for high- and low-level features. In the first row, two images captured at different time points, namely, the ground-truth and the result of the LSS-Net, are shown from left to right. In the second row, the first and second images are the attention maps of a local similarity attention module for low- and high-level features, respectively. A thicker red indicates an area where change has occurred, whereas a thicker blue indicates an area where no change has occurred. As shown in Fig. 7, the low-level local similarity attention module detects changes in low-level features, such as edges and dots. Whereas the high-level local similarity attention module detects changes in semantic information, such as the location information of the changed buildings in both input images. In other words, the high-level local similarity attention module detects the position of a roughly changed object. In contrast, the low-level local similarity attention module infers the shape of a changed object. In the second row, the third and fourth images are the overlaid images of attention maps of low- and high-level features and an image_now, respectively. These figures show what was detected by the local similarity attention module of each level feature.

The decoder combines these two pieces of information. From the results, we can see that our method produces a shape that is more similar to that of the ground-truth when compared to those of the other methods, as shown in Fig. 8. From this analysis, we verified that it is extremely effective to calculate the similarity of the change detection by dividing the low- and high-level characteristics.

We also compared the performance of networks using only low-level or high-level local similarity attention modules with that of the LSS-Net to ensure the effects of our proposed local similarity attention module on the LSS-Net. The F1 score of networks with the low-level local similarity attention module was 0.7332, whereas that of the network with the high-level local similarity attention module was 0.6798. The F1 scores of networks with only a low- or high-level local similarity attention module were lower than those of the LSS-Net. Through this experiment, we could ensure that the decoder structure capable of combining low- and high-level features improves the performance of the LSS-Net.

C. Comparison With State-of-The-Art Methods

In this subsection, we evaluate the performance of our proposed network. We used the following evaluation metrics to evaluate the performance of the LSS-Net: Accuracy, F1 score, precision, recall (sensitivity), and Cohen's kappa coefficient. The accuracy, F1 score, precision, and recall are the most commonly used metrics for binary segmentation problems, and the kappa coefficient was used as a comprehensive metric. Because the kappa coefficient is robust to imbalanced data, the kappa coefficient is the most commonly used metric for change detection when dealing with unbalanced data. Because change detection is a very imbalanced task, we excluded the accuracy

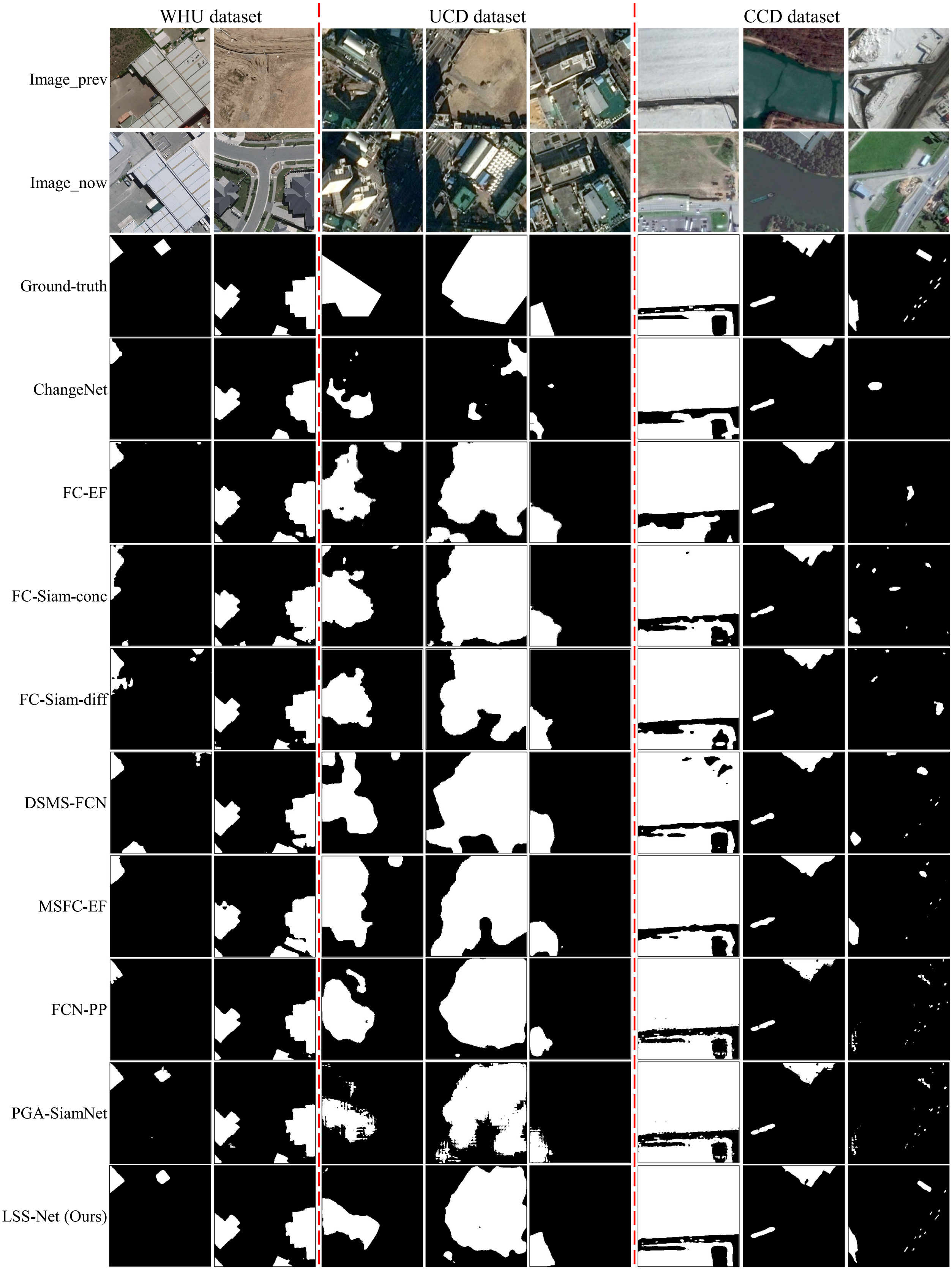


Fig. 8. Qualitative comparison for WHU, UCD, and CDD datasets. The two columns on the left are from the WHU dataset, the middle three columns are from the UCD dataset, and the three columns on the right are from the CDD dataset.

TABLE IV
QUANTITATIVE COMPARISON OF CHANGE DETECTION METHODS ON CDD [28], WHU [29], AND UCD DATASET

Dataset	Method	F1 score	Precision	Recall (Sensitivity)	Kappa
CDD dataset [28]	ChangeNet [35]	0.9026	0.9216	0.8844	0.8899
	FC-EF [15]	0.8773	0.9061	0.8503	0.8614
	FC-Siam-conc [15]	0.8943	0.8973	0.8913	0.8802
	FC-Siam-diff [15]	0.8664	0.9042	0.8316	0.8494
	DSMS-FCN [33]	0.9144	0.9165	0.9124	0.9030
	MSFC-EF [33]	0.9065	0.9030	0.9102	0.8940
	FCN-PP [18]	0.9213	0.9417	0.9019	0.9111
	PGA-SiamNet [19]	0.9347	0.9657	0.9056	0.9263
	LSS-Net_VGG (Ours)	**0.9652**	**0.9706**	**0.9599**	**0.9606**
	LSS-Net (Ours)	0.9630	0.9674	0.9587	0.9581
WHU dataset [29]	ChangeNet [35]	0.8945	0.9323	0.8582	0.8904
	FC-EF [15]	0.8496	0.8296	0.8707	0.8428
	FC-Siam-conc [15]	0.7663	0.6596	0.9141	0.7550
	FC-Siam-diff [15]	0.8174	0.7363	0.9185	0.8090
	DSMS-FCN [33]	0.7772	0.6704	0.9179	0.7665
	MSFC-EF [33]	0.8546	0.8032	0.9131	0.8482
	FCN-PP [18]	0.9120	0.9412	0.8839	0.9085
	PGA-SiamNet [19]	0.9289	0.9279	0.9299	0.9260
	LSS-Net_VGG (Ours)	0.9345	0.9258	**0.9435**	0.9318
	LSS-Net (Ours)	**0.9377**	**0.9418**	0.9336	**0.9351**
UCD dataset	ChangeNet [35]	0.5489	0.8172	0.4133	0.5399
	FC-EF [15]	0.6438	0.5945	0.7020	0.6242
	FC-Siam-conc [15]	0.6722	0.5458	**0.8745**	0.6515
	FC-Siam-diff [15]	0.6246	0.5227	0.7759	0.6016
	DSMS-FCN [33]	0.6295	0.4959	0.8617	0.6054
	MSFC-EF [33]	0.6429	0.5222	0.8361	0.6204
	FCN-PP [18]	0.7035	0.8054	0.6245	0.6905
	PGA-SiamNet [19]	0.7227	0.8274	0.6415	0.7105
	LSS-Net_VGG (Ours)	0.7405	**0.8546**	0.6221	0.7082
	LSS-Net (Ours)	**0.7751**	0.8409	0.7188	**0.7646**

Best performance is in **bold**, and the second-best performance is underlined

metric. The equations for all evaluation metrics are as follows:

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{5}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

$$p_c = \frac{(TP{+}FP)(TP{+}FN) + (FP + TN)(FN{+}TN)}{(TP{+}FP{+}FN{+}TN)^2} \tag{8}$$

$$\text{Kappa} = \frac{\text{Accuracy} - p_c}{1 - p_c} \tag{9}$$

where *TP*, *TN*, *FP*, and *FN* are the true positive, true negative, false positive, and false negative, respectively.

We compared our models with several state-of-the-art approaches for change detection, namely, FC-EF, FC-Siam-conc, FC-Siam-diff [15], ChangeNet [35], MSFC-EF, DSMS-FCN [33], FCN-PP [29], and PGA-SiamNet [19]. FC-Siam-conc, FC-Siam-diff, ChangeNet, DSMS-FCN, and PGA-SiamNet are Siamese network architectures, which consist of two identical networks that share the weights of the convolution layers. FC-EF, MSFC-EF, and FCN-PP are early fusion architectures that concatenate the two input images before passing them through the network. For a fair comparison, we also compared a version of the LSS-Net using VGG-Net as the backbone with other methods. This version of the LSS-Net is denoted as LSS-Net_VGG.

a) Quantitative Comparison: We first quantitatively compared our models with other state-of-the-art methods in terms of accuracy, F1 score, precision, recall, and Cohen's kappa coefficient (Kappa) [36]. A higher kappa value indicates that the image is more similar to the ground-truth image. All performance values for the other methods were obtained through the implementation of networks in the same environments. We trained and tested all models on the same datasets, namely, the CDD, WHU, and UCD training and test datasets.

Table IV shows quantitative comparisons of the CDD, WHU, and UCD datasets. It can be seen that the proposed network, LSS-Net, outperforms the state-of-the-art models in terms of accuracy, F1 score, precision, and kappa coefficient. Thus, these results allow us to validate the effectiveness of both the local similarity attention module and the change attention map-based content loss function. LSS-Net produces higher kappa values for the three datasets than other models. Because the kappa coefficient indicates agreement with the ground-truth, it demonstrates that the results of our model are the most similar to the ground-truth, ensuring that the LSS-Net_VGG outperforms the state-of-the-art methods. The recall values of the LSS-Net and LSS-Net_VGG are slightly lower than those of other methods for the UCD dataset. This may be caused by the complexity of the UCD dataset, which shows a very complex urban area compared to the other two datasets. The changed areas in the UCD dataset occupy relatively small regions compared to the overall area. The LSS-Net and LSS-Net_VGG were designed to predict a better shape of the changed areas in addition to the F1 score by using a local similarity module. The degraded recall values of the LSS-Net and LSS-Net_VGG due to these inherent network characteristics need to be further elucidated in future studies.

b) Qualitative Comparison: Fig. 8 shows a qualitative comparison between our model and the other state-of-the-art models on the WHU, UCD, and CDD datasets. This figure shows that our model produces more accurate results than other models. The images in the first and second rows are input images, and those in the third row are ground-truths. The other images are the results of the state-of-the-art models and LSS-Net. The first two columns indicate the results for the WHU dataset, the next three columns show the results for the UCD dataset, and the last three columns show the results for the CDD dataset. According to the first, second, third, and fourth columns, our models produced much better shapes than other methods. In the first column, the LSS-Net can detect a new and a disappearing building, whereas most of the methods except PGA-SiamNet cannot detect a disappearing building. In particular, according to the third column, the LSS-Net during change detection is more robust to shadows as compared to the other models. According to the fourth column, although our result is slightly different from the corresponding ground-truth, the LSS-Net produces a better shape of a detected changed area for a newly constructed building than the other models. In the sixth and eighth columns, the results show that the LSS-Net offers excellent performance in change detection despite the seasonal changes. In the seventh column, our results are also the most similar to the ground-truth. Therefore, these results demonstrate that our model outperforms other state-of-the-art models in terms of the predicted shape and accuracy for the CDD dataset.

IV. Discussion

In this study, we proposed an LSS-Net and a change attention map-based content loss function to improve the change detection in remote sensing images. The performance of our proposed method, LSS-Net, was compared with those of other state-of-the-art methods for change detection with two public datasets, namely, the CDD and WHU datasets, and our UCD dataset. The experimental results reveal that the LSS-Net outperforms other state-of-the-art models in the task of change detection. In the evaluation using two public datasets, LSS-Net exhibited the highest F1 scores of 0.9630 and 0.9377 on the CDD and WHU datasets, respectively. In addition, it offered a better performance than other models with regard to change detection on a complex urban dataset (UCD dataset). It provided the highest F1 score of 0.7751, which is 0.05 to 0.23, which is higher than those of other methods. It is highly important to detect the exact shapes of changed objects during change detection in urban complex cities because buildings are small and dense, causing false-positive and false-negative rates to rapidly increase compared to the true-positive rate when the building shape is not properly detected during change detection. In particular, compared to other methods, LSS-Net offers superior performance for the prediction of sophisticated shapes of changed objects from complex satellite images (Fig. 8).

Moreover, we analyzed the local similarity attention module of the LSS-Net to examine its functionality in change detection (Fig. 7). During the analysis, the attention module was determined to be simpler and more effective than the conventional attention module of PGA-SiamNet [19]. The local similarity attention module divides low- and high-level features into two stages to obtain output feature maps, whereas the attention module of PGA-SiamNet hierarchically stacks decoders to obtain the output feature maps. Therefore, the local similarity attention module is simpler but offers a better performance than that of the attention module of PGA-SiamNet for change detection, as shown in Table IV.

For a fair comparison, the performance of the LSS-Net was also compared with those of other models in the change detection task when VGG-Net was utilized as an encoder of the LSS-Net. The LSS-Net based on a VGG-Net encoder also exhibited a better performance than other models in change detection. Interestingly, the LSS-Net based on the VGG-Net encoder exhibited the highest performance on the CDD dataset (Table IV). However, the performance of the LSS-Net based on the VGG-NET encoder was similar to that of the LSS-Net.

In addition, we verified the effectiveness of the proposed change attention map-based content loss function (Table II, Table III, and Table IV). The proposed loss function improved the F1 score of our model by 0.7751, which is 0.0246 higher than that of our model trained using only a BCE loss function. We applied change attention map-based content loss to other methods. As a result, the performance of the methods was improved from 0.0020 to 0.0394 in terms of the F1 score. Unlike the BCE loss function, our proposed change attention map-based content loss function can use the content information on input images. Thus, it can improve the performance of the LSS-Net in change detection.

V. Conclusion

We demonstrated that our proposed LSS-Net outperforms other state-of-the-art models in change detection from remote sensing images. In addition, we ensured the superiority of the proposed change attention map-based content loss function compared to other loss functions. Moreover, the attention module and decoder structure of the LSS-Net were found to be much simpler and more effective than the other methods such as PGA-SiamNet. The results presented in this study demonstrate that the LSS-Net optimized with a change attention map-based content loss function offers a state-of-the-art performance in the change detection of remote sensing images, thus suggesting its usefulness as a tool for various remote sensing applications, such as the monitoring of urban expansion and illegal building construction. In this study, LSS-Net with a change attention map-based content loss function was applied only to the binary change detection of remote sensing images. However, multiclass change-detection tasks are also important in remote sensing, and the LSS-Net can be beneficial for such tasks. Such challenges can be an area of focus in future research.

REFERENCES

[1] P. P. Singh and R. Garg, "Automatic road extraction from high resolution satellite image using adaptive global thresholding and morphological operations," *J. Indian Soc. Remote Sens.*, vol. 41, no. 3, pp. 631–640, 2013.

[2] J. H. Kim *et al.*, "Objects segmentation from high-resolution aerial images using U-Net with pyramid pooling layers," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 1, pp. 115–119, Jan. 2019.

[3] Y. Hua, L. Mou, and X. X. Zhu, "Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 149, pp. 188–199, 2019.

[4] Y. Na, J. H. Kim, K. Lee, J. Park, J. Y. Hwang, and J. P. Choi, "Domain adaptive transfer attack-based segmentation networks for building extraction from aerial images," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.3010055.

[5] S. Liu, Q. Shi, and L. Zhang, "Few-shot hyperspectral image classification with unknown classes using multitask deep learning," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.3018879.

[6] Q. Shi *et al.*, "Domain adaption for fine-grained urban village extraction from satellite images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1430–1434, Aug. 2020.

[7] A. Singh, "Review article digital change detection techniques using remotely-sensed data," *Int. J. Remote Sens.*, vol. 10, no. 6, pp. 989–1003, 1989.

[8] P. Coppin, I. Jonckheere, K. Nackaerts, B. Muys, and E. Lambin, "Review articledigital change detection methods in ecosystem monitoring: A review," *Int. J. Remote Sens.*, vol. 25, no. 9, pp. 1565–1596, 2004.

[9] S. Nghiem, K. Steffen, R. Kwok, and W. Tsai, "Detection of snowmelt regions on the Greenland ice sheet using diurnal backscatter change," *J. Glaciology*, vol. 47, no. 159, pp. 539–547, 2001.

[10] C. Munyati, "Wetland change detection on the Kafue flats, Zambia, by classification of a multitemporal remote sensing image dataset," *Int. J. Remote Sens.*, vol. 21, no. 9, pp. 1787–1806, 2000.

[11] C. Atzberger, "Advances in remote sensing of agriculture: Context description, existing operational monitoring systems and major information needs," *Remote Sens.*, vol. 5, no. 2, pp. 949–981, 2013.

[12] L.-C. Chen and L.-J. Lin, "Detection of building changes from aerial images and light detection and ranging (Lidar) data," *J. Appl. Remote Sens.*, vol. 4, no. 1, 2010, Art. no. 041870.

[13] N. Bourdis, D. Marraud, and H. Sahbi, "Constrained optical flow for aerial image change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2011, pp. 4176–4179.

[14] C. Benedek and T. Szirányi, "Change detection in optical aerial images by a multilayer conditional mixed Markov model," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 10, pp. 3416–3430, Oct. 2009.

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

[16] 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.

[17] 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, Mar. 2018.

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

[19] 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, 2020, Art. no. 484.

[20] J. Zhang, T. Li, X. Lu, and Z. Cheng, "Semantic classification of high-resolution remote-sensing images based on mid-level features," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 6, pp. 2343–2353, Jun. 2016.

[21] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[22] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3085–3094.

[23] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1874–1883.

[24] W. Cai and Z. Wei, "Remote sensing image classification based on a cross-attention mechanism and graph convolution," *IEEE Geosci. Remote Sens. Lett.*, early access, Oct. 1, 2020, doi: 10.1109/LGRS.2020.3026587.

[25] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.

[27] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[28] 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. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. XLII-2, pp. 565–71, 2018.

[29] 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.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[31] A. Paszke *et al.*, "Automatic differentiation in PyTorch," in *Proc. 31st Conf. Neural Inf. Process. Syst., Autodiff Workshop*, Dec. 2017, pp. 1–4.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

[33] H. Chen, C. Wu, B. Du, and L. Zhang, "Deep Siamese multi-scale convolutional network for change detection in multi-temporal VHR images," in *Proc. 10th Int. Workshop Anal. Multitemporal Remote Sens. Images*, 2019, pp. 1–4.

[34] A. Dosovitskiy *et al.*, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2758–2766.

[35] A. Varghese, J. Gubbi, A. Ramaswamy, and P. Balamuralidhar, "ChangeNet: A deep learning architecture for visual change detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp 129–145.

[36] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.

Haeyun Lee (Student Member, IEEE) was born in Iksan, South Korea. He received the B.S. degree in mathematics from Chonbuk National University, Jeonju, South Korea, in 2016, and the M.S. degree in information and communication engineering from Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu, South Korea, in 2018, where he is currently working toward the Ph.D. degree in information and communication engineering.

His current research interests include the image restoration and medical image analysis with deep learning.

Kyungsu Lee (Student Member, IEEE) received the B.S. degree in computer science from Handong Global University, Pohang, South Korea, in 2018. He is currently working toward the Ph.D. degree in information and communication engineering from Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu, Korea.

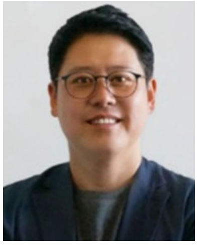
His research interest includes deep-learning in the biomedical application and remote sensing.

Jun Hee Kim (Member, IEEE) received B.S. degree in electrical engineering from Chungnam National University, Daejeon, South Korea, in 2014, and the M.S. and Ph.D. degrees in information and communication engineering from Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu, South Korea, in 2016 and 2020, respectively.

Since 2020, he has been a Senior Researcher with the Agency for Defense Development, Daejeon, South Korea. His research interests include deep learning for satellite image segmentation, resource allocation in RF-powered communication networks, and cooperative spectrum sensing in cognitive radio networks.

Younghwan Na (Student Member, IEEE) received the M.S. degree in information and communication engineering from Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu, Korea, in 2021.

His research interests include computer vision and machine learning.



Juhum Park (Member, IEEE) received the B.S degree in material science and engineering from Korea University, Seoul, South Korea.

He worked with LG Electronics at HQ Korea and France and Czech Republic, responsible for product and business strategy. He founded Dabeeo, Inc., Seoul, South Korea, in 2012, which is working on Geospatial data business. Since 2016, he has been developing deep learning technology for geospatial data generation which is based on image segmentation and image processing.

Jihwan P. Choi (Senior Member, IEEE) received the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA.

He is currently an Associate Professor with the Department of Aerospace Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea. He was with Marvell Semiconductor Inc., Santa Clara, CA, USA and with the Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu, Korea. His research interests include aerospace and wireless communications, and the applications of machine learning and deep learning.

Dr. Choi is an Associate Editor for the IEEE TRANSACTIONS ON AEROSPACE AND ELECTRONIC SYSTEMS and IEEE ACCESS, and an Editorial Board Member for *Remote Sensing*.

Jae Youn Hwang (Member, IEEE) received the Ph.D. degree in biomedical engineering from the University of Southern California, Los Angeles, CA, USA, in 2009.

He is currently an Associate Professor with the Department of Information and Communication Engineering, Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu, South Korea. His current research interests include the development of an intelligent multimodality imaging systems, based on high frequency ultrasound and optical techniques, and the applications of machine learning in medicine and remote sensing.