International Journal of Remote Sensing

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tres20

EMS-CDNet: an efficient multi-scale-fusion change detection network for very high-resolution remote sensing images

Zhi Zheng, Yi Wan, Yongjun Zhang, Kun Yang, Rang Xiao, Chao Lin, Qiong Wu & Daifeng Peng

To cite this article: Zhi Zheng, Yi Wan, Yongjun Zhang, Kun Yang, Rang Xiao, Chao Lin, Qiong Wu & Daifeng Peng (2022) EMS-CDNet: an efficient multi-scale-fusion change detection network for very high-resolution remote sensing images, International Journal of Remote Sensing, 43:14, 5252-5279, DOI: 10.1080/01431161.2022.2131479

To link to this article: https://doi.org/10.1080/01431161.2022.2131479

INTERNATIONAL JOURNAL OF REMOTE SENSING
2022, VOL. 43, NO. 14, 5252–5279
https://doi.org/10.1080/01431161.2022.2131479

EMS-CDNet: an efficient multi-scale-fusion change detection network for very high-resolution remote sensing images

Zhi Zheng[a], Yi Wan[a], Yongjun Zhang[a], Kun Yang[b], Rang Xiao[c], Chao Lin[d], Qiong Wu[a] and Daifeng Peng[e]

[a]School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China; [b]Guizhou Basic Geographic Information Center, Guiyang, China; [c]Guizhou Tuzhi Information Technology, Guiyang, China; [d]Land resource and information center of Guangdong province, Guangzhou, China; [e]School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing, China

ABSTRACT

Remote sensing image change detection (RSICD) is an essential measure for monitoring the earth's surface changes. In recent years, the explosive growth of very high-resolution (VHR) satellite sensors and the booming innovations in deep learning technology have significantly boosted RSICD development. However, most of the current RSICD models focus on locating accurate change areas while ignoring the efficiency of their method, which limits the practical application of RSICD models, especially for large-scale and emergency RSICD tasks. In this paper, we propose an Efficient Multi-scale-fusion Change Detection Network (EMS-CDNet) for bi-temporal RSICD tasks. Our EMS-CDNet pays more attention to the model's inference speed and the accuracy-efficiency trade-off rather than only pursuing detection accuracy. We designed a multi-scale fusion module for EMS-CDNet, which adopts multi-scale and multi-branch operations to extract multi-scale features simultaneously and aggregate features at different feature levels. In addition to EMS-CDNet's ability to achieve sufficient feature extraction, the multi-scale image input within the designed module alleviates the influence of image registration errors in practical applications, thereby strengthening EMS-CDNet's value for practical RSICD tasks. We also integrated a novel partition unit in EMS-CDNet to lighten the model while maintaining the detection ability of small targets, thus shortening its processing time without a severe accuracy decrease. We conducted experiments on two state-of-the-art (SOTA) public RSICD datasets and our own collected dataset. The public datasets were utilized to comparatively measure the overall accuracy and efficiency measurement of EMS-CDNet, and the dataset of images we collected was used to observe EMS-CDNet's performance under the influence of image registration errors. Our experimental results show that EMS-CDNet achieved a better accuracy-efficiency trade-off than the SOTA public datasets methods. For example, EMS-CDNet reduced the inference time by about 33%

ARTICLE HISTORY
Received 28 January 2022
Accepted 27 September 2022

KEYWORDS
Change detection (CD); Multi; scale; fusion module; partition unit; accuracy; efficiency trade; off

CONTACT Yi Wan yi.wan@whu.edu.cn School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China; Yongjun Zhang zhangyj@whu.edu.cn

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

while maintaining identical detection accuracy to CLNet (the optimal method among the comparison methods). Furthermore, EMS-CDNet achieved higher accuracy on our collected dataset, with an F1 of 74% and mIoU of 0.806, demonstrating its robustness to image registration errors and showing its value for practical RSICD applications.

1. Introduction

Remote sensing image change detection (RSICD) is the current technology for monitoring the dynamic evolution of the earth's surface and detecting changes in natural resources or artificial structures in the same geographical area. RSICD is one of the current research hotspots in the remote sensing community (Zhu et al. 2017; Lv et al. 2022b; Wen et al. 2021) and has been broadly applied in various tasks, such as land use monitoring (Jin et al. 2013), urban expansion (Lu et al. 2010), land cover mapping (Yang et al. 2003), and disaster damage investigation (Wang and Xu 2010).

The RSICD methods created to date mainly utilize hand-draft feature extractors to analyze bi-temporal images and detect changes. Since the resolution of the acquired images was relatively lower in the early stages of RSICD, researchers treated the pixel as the primary processing unit and developed extensive pixel-level RSICD methods. For example, mathematical and statistical techniques, such as principal component analysis (PCA), slow feature analysis (SFA), and canonical correlation analysis (CCA), were embedded into the RSICD framework and have proven to be effective (Celik 2009; Wu et al. 2017; Nielsen 2007). However, once the need was recognized for setting predetermined thresholds for different RSICD tasks and geographical conditions, these methods became unreliable in complex observation cases. At the same time, more complicated algorithms were being introduced to improve detection confidence and accuracy, such as the image texture analysis and the conditional random field-based methods (Erener and Düzgün 2009; Lv et al. 2016). With the resolution improvement of remote sensing images, however, side effects (such as mixed pixels and insufficient feature representation problems) became more severe and thus hindered RSICD accuracy (Dalla Mura et al. 2008; Chen et al. 2012).

To overcome the limitations of the pixel-level methods, numerous object-level RSICD methods have been introduced in the last few decades (Zheng et al. 2021; Lv et al. 2022a). Object-level methods first segment the image pairs into multiple homogeneous segments by interpreting the spectral contextual and geometric information and then taking them as input to determine the changed areas (Bock et al. 2005; Hussain et al. 2013). Lefebvre, Corpetti, and Hubert-Moy (2008) proposed a geometric index for change determination, while Diego et al. (2012) utilized spectral intensity to detect changes. Zhou, Troy, and Grove (2008) combined the object size, shape, and adjacency as their change detection rules. Zhang, Peng, and Huang (2017) further utilized a multi-scale uncertainty analysis strategy to take advantage of contextual information. In addition to the foregoing methods that use geometric or spectral information, other methods were developed that integrate multi-source data or multiple types of image features for comprehensive analysis. For example, Tomowski, Ehlers, and Klonus (2011) compared

the texture and the spectral information of different objects and Gamanya, De Maeyer, and De Dapper (2009) combined image information with GIS layers and multi-temporal data to distinguish the actual changes. The object-level RSICD methods are effective for object representation and can significantly enhance the accuracy of RSICD; however, their performance significantly relies on the prefix segmentation algorithms.

Deep learning techniques are showing immense potential for image understanding (Zhu et al. 2017; Wang et al. 2018), and their effectiveness for RSICD tasks has been broadly investigated in recent years (Peng, Zhang, and Guan 2019; Zheng et al. 2021; Lv et al. 2022c). Depending on their adopted network types, the deep-learning-based RSICD methods can be divided into four categories: 1) Convolutional Neural Networks (CNNs)-based RSICD methods; 2) Recurrent Neural Networks (RNNs)-based RSICD methods; 3) Generative Adversarial Networks (GANs)-based RSICD methods, and 4) hybrid networks-based RSICD methods. The CNN-based RSICD methods regard RSICD tasks as image classification problems and use CNNs as solvers (Gong et al. 2015). Wang et al. (2018) trained a CNN to extract useful information for multispectral change detection tasks. Ji et al. (2019) considered a building's change detection as an instance segmentation task and trained a CNN to extract features and used Fast-RCNN (Girshick 2015) to determine the change areas. More advanced methods, such as the attention-based and pyramid-based methods, further considered the spatial location information and designed specific feature extractors to enhance the model's detection accuracy (Zhang et al. 2020; Lin et al. 2017). For the RNN-based RSICD methods, researchers first transformed bi-/multi-temporal image patches into time-sequential data to meet the input requirements of RNNs and then utilized their powerful sequential data processing ability to facilitate detection accuracy. Lyu, Lu, and Mou (2016) used long-short-term memory (LSTM) to deal with multispectral and hyperspectral change detection. They utilized a core memory cell to learn the change rule and three gates to control the model's input, output, and update parameters. Numerous training samples are required to train RSICD networks (Lv et al. 2020). Thus, GAN-based RSICD methods were introduced to ease the high requirements of labeled samples. Gong et al. (2019) designed a GAN to generate unlabeled and new fake data, while Huang, Zhang, and Wang (2020) designed a special GAN to clean up training samples collected under noise conditions. Saha, Bovolo, and Bruzzone (2019) transformed image pairs into the same domain and detected the changed areas through deep feature change vector analysis. In addition to the three types of methods, researchers also have investigated the potential of combining different types of networks for better RSICD results. For example, Mou, Bruzzone, and Zhu (2018) designed a recurrent CNN to train specific temporal features and concluded that the hybrid network could obtain better results than using a single CNN or RNN. Song et al. (2018) proposed a 3DCNN and a ConvLSTM to exploit the spatial-spectral-temporal information of multispectral images and thus achieve promising detection accuracy. Furthermore, Maria et al. (2019) combined a CNN and LSTM that achieved about 95% overall accuracy for the task of urban CD.

CNN-based methods are the most investigated among the four types of deep-learning-based methods. Due to the high correlation between RSICD and semantic segmentation tasks, various typical semantic segmentation models, such as FCN (Long, J., E. Shelhamer, and T. Darrell (2015)), Unet (Ronneberger, Fischer, and Brox 2015), and their variations, were extended into the RSICD field. Daudt et al. (2018)

proposed three FCN-based networks to concatenate the input image pairs to learn the joint features and named them as follows: 1) the fully convolutional-early fusion network (FC-EF), 2) the fully convolutional Siamese-concatenation network (FC-Siam-conc), and 3) the fully convolutional Siamese difference network (FC-Siam-diff). FC-Siam-conc and FC-Siam-diff are from FC-EF in that they take two weight-share branches as feature extractors to extract the independent image features. Then, the extracted features are fused to distinguish the changed areas. The difference between the two models is called the feature fusion approach, where the former adopted the element-minus strategy while the latter took the element-adds strategy. Peng, Zhang, and Guan (2019) aggregated features extracted from different scales of the Unet++ (Zhou et al. 2018) and applied multiple side-output fusion strategies to generate better change maps. Their method combined the advantages of multi-scale features and multi-level content information and thus was effective on satellite images.

Despite performing well in some cases, the Unet-series RSICD methods still have shortcomings. Among these methods, some tend to use lightweight models, such as FC-EF and FC-Siam-conc. As a result, the inference time is shortened, whereby the detailed changes in the scenario may not be detected and its detection accuracy may be limited. The other methods pursue higher prediction accuracy rather than the model's efficiency. For example, the method of Peng, Zhang, and Guan (2019) achieves higher detection accuracy while its adopted deep supervision strategy and multi-level fusion structure are quite time-consuming. Given these issues, achieving a better accuracy-efficiency trade-off is essential for the RSICD tasks, especially facing large-scale and emergency demands (i.e. how to lighten the RSICD models while retaining the detection accuracy simultaneously). Although numerous attempts have been made to lighten models (Mehta et al. 2018; Howard et al. 2017; Sandler et al. 2018; Zhang et al. 2018), directly transferring them for the RSICD tasks always resulted in a severe accuracy decrease. In addition, most of the existing RSICD models achieved high accuracy on the given datasets, but they did not perform well in practical applications. One of the main reasons for this poor performance is that they are trained on artificially cleaned-up datasets, and thus some obstacles in practical applications, such as image registration errors, are not considered during the training process. Therefore, in this paper, we propose a novel EMS-CDNet method to address the accuracy-efficiency trade-off problem and improve the model's robustness to image registration errors in practical applications. The contributions of this paper are as follows:

(1) A novel multi-scale fusion (MSF) strategy to extract the primary features sufficiently and improve the model's detection performance. Since our MSF strategy enables multi-scale images for network input, it also can alleviate the side-effects of image registration errors and improves the model's value in practical applications.
(2) A novel partition unit to lighten the model derived from the group convolution (Cohen et al., 2016). Our partition unit lightens the model and suppresses the accuracy decrease problem of group convolution, thereby achieving a better accuracy-efficiency trade-off than the comparison methods.
(3) Our proposed EMS-CDNet is an integration of the above MSF module and partition unit. Our experiments on two public RSICD datasets and our own collected dataset demonstrated EMS-CDNet's superior performance. On the public datasets, it

achieved results similar to SOTA methods while shortening the processing time. EMS-CDNet also performed best on our collected dataset, benefitting from its robustness to image registration errors.

The remainder of this paper is organized as follows. Section 2 presents the details of the proposed modules and the constructed network. Section 3 presents our investigation of the effectiveness of EMS-CDNet and our analysis based on a comparison with the results of state-of-the-art (SOTA) FCN- and Unet-series RSICD methods. Section 4 discusses the performance and limitations of EMS-CDNet in detail. Finally, Section 5 presents our concluding remarks.

2. Materials and methods

In this section, we introduce the multi-scale-fusion input strategy, illustrate our theory behind the partition unit, discuss the details of the structure of our proposed EMS-CDNet, and introduce our loss function.

2.1. Multi-scale fusion (MSF) strategy

Based on the high correlation between the RSICD and semantic segmentation tasks, researchers have proposed many effective change detection methods that refer to semantic segmentation models. Numerous multi-scale feature extraction modules also have been proposed to improve the model's feature representation capability and to make it more effective for accuracy improvements. However, these existing modules still have some limitations. First, most of the multi-scale modules only bridge the adjacent features and thus ignore the spatial correlations and detection of the detailed structures. Second, the proposed complex feature extraction strategies complicate the models and result in a lengthy processing time. In addition, since the bi-temporal images for RSICD were captured with different sensors and taken from different viewing perspectives, they inevitably have image registration errors. Most of these methods neglect this problem, and that is why they perform well on strictly registered images while failing in practical applications. We, therefore, propose a multi-scale fusion (MSF) strategy in this paper as the network input to achieve sufficient feature extraction and alleviate the side-effects of image registration errors, thus strengthening the model's performance in practical applications. As shown in Figure 1, the MSF module has four branches. The first two branches utilize dilation convolutions with different strides and rates to extract multi-scale features. Then, the extracted feature maps are down-sampled utilizing max-pooling operations. The last two branches first resize the raw images to their 1/8 and 1/16 counterparts and then transform the resized images into the feature domain by 1×1 convolutions. Since the last two branches down-sample the raw images and are embedded into the network in their original scale, they can somewhat alleviate the image registration errors. Our MSF strategy enables multi-scale network inputs, which are 1/2, 1/4, 1/8, and 1/16 of the raw images' sizes. Then, a two-branches module is applied to the first branch, making it possible to concatenate the features from the first two branches for further information aggregation.

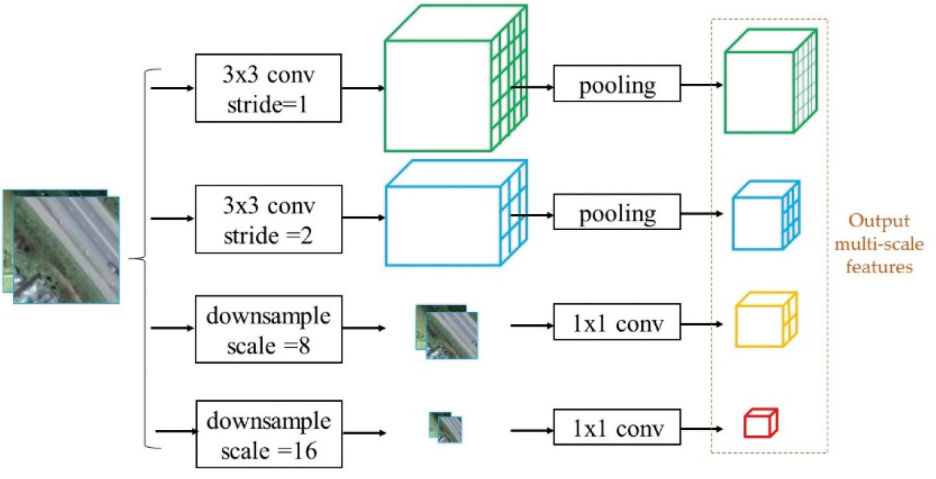


Figure 1. Multi-scale-fusion (MSF) input.

2.2. Network lighten operations

Most of the existing models designed for complex structures ensure high detection accuracy while sacrificing their efficiency. However, the efficiency of RSICD models is essential for practical applications, especially for large-scale and emergency cases. There are some strategies, such as group convolution (Cohen and Welling 2016) and separable convolution (Chollet 2017), that have been adopted to lighten the models; however, they cannot achieve high accuracy in practical applications. Therefore, we designed a novel partition unit to lighten the model and maintain the model's accuracy performance.

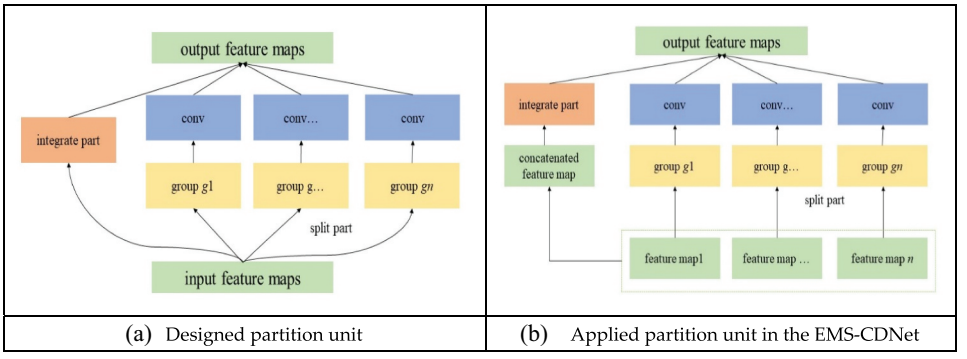
Our partition unit is derived from group convolution (Cohen and Welling 2016), which splits input feature maps into several equal groups according to the channel dimension. The structure of group convolution is as follows. Suppose the input and the output feature maps have the same size and their channels are respectively C_{in} and C_{out}, the kernel size of the operation convolutions is k, and the group numbers are g. Then, the regular and the group convolution parameters can be estimated with Equation (1) and (2), respectively. It can be seen that the parameters of group convolution are only $1/g$ of the regular convolution. Therefore, the group convolution process can significantly decrease the model parameters with proper group numbers.

$$Para_{regular} = C_{in} \times C_{out} \times k^2 \tag{1}$$

$$Para_{group} = g \times \left(\frac{C_{in}}{g} \times \frac{C_{out}}{g} \times k^2 \right) = \frac{1}{g} \times Para_{regular} \tag{2}$$

The group operation divides the feature maps into independent groups so the contextual information among different feature groups cannot be fully exchanged. Therefore, although the models can be lightened by group convolution, the accuracy decreases.

The goal of our proposed EMS-CDNet is to lighten the model while maintaining detection accuracy. Therefore, we adopted a variation of group convolution, called the partition unit. As shown in Figure 2(a), the partition unit first divides the concatenate



(a) Designed partition unit (b) Applied partition unit in the EMS-CDNet

Figure 2. Illustration of the partition unit.

feature maps into two independent parts: the split part and the integrated part. The split part is then divided into *g* groups in the same manner as group convolution. The split part helps to lighten the model. The integrated part aggregates the remaining feature maps and embeds them into the output feature maps. The integrated part maintains a portion of the semantic features and thus alleviates the accuracy problem. We calculate the partition unit's parameters in Equation (3) to illustrate its effectiveness intuitively.

$$Para_{partition} = \left(C_{in} \times \frac{C_{out}}{2} + \sum_{i=1}^{g} C_{in\_i} \times C_{out\_i} \right) \times k^2 \tag{3}$$

Where $C_{in_i}$and C_{out_i} represent the input and output channels of each group in the split part.

Our MSF strategy can generate multiple feature maps on different scales. Therefore, we adjusted the input of the designed partition unit in the EMS-CDNet to satisfy the requirements of the MSF module. Figure 2(b) illustrates the process adopted in EMS-CDNet for integrating the partition units with the feature.

2.3. Overall network structure

Thus far, by integrating the MSF input with the partition unit, we created our proposed EMS-CDNet, as depicted in Figure 3. The image pairs are first concatenated along the channel dimension to meet the requirements of the network input. With the application of the MSF input, the preliminary features are extracted, whose sizes are 1/2, 1/4, 1/8, and 1/16 of the network input (marked as *L1-0*, *L2-0*, *L3–0*, and *L4-0* in Figure 3). Then, a two-branch module is attached to *L1-0* to extract the higher-level features (marked as *L2-1*, and *L2-2* in Figure 3). To enlarge the receptive field of the extracted features, we take dilation convolutions with different rates in the two-branch module as basic blocks. In this stage, the feature maps *L2–0, L2–1*, and *L2-2* have the same size, and they are concatenated to strengthen the model's representation ability. To extract deeper image features and lighten the proposed EMS-CDNet simultaneously, the first partition unit is applied, which takes the *L2-0*, *L2-1*, *L2–2*, and *L2-c* as input, and outputs feature maps *L3-1*, *L3–2, L3–3*, and *L3–4*. The second partition unit takes the third branch of the MSF input and the output of the first partition unit as input and outputs feature maps *L4-1* through *L4–6*. The

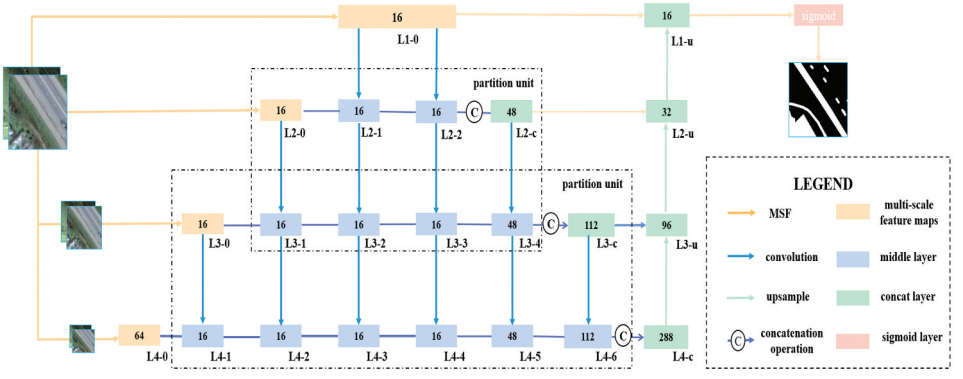


Figure 3. The overall structure of the EMS-CDNet. (Values in each layer indicate the channel numbers)

detailed implementation of the partition unit is shown in the black dashed boxes in Figure 3. Since the partition units are adopted, the channels of the deeper feature maps are not expanded; thus, feature maps *L2–1*, *L2–2*, *L3-1* through *L3–3*, and *L4-1* through *L4-4* have the same channels as the preliminary features. The decoder part of the EMS-CDNet is similar to the typical Unet structure, where the feature maps come from the encoder part and the up-sampling operations are concatenated to suppress detail losses and recover accurate boundaries. Finally, a 3×3 convolution with the sigmoid activation function is adopted to generate the predicted change maps.

2.4. Loss function

For the RSICD tasks, binary cross entropy (BCE) loss is widely used. In EMS-CDNet, we also use it as a part of the loss function. The BCE loss function is defined as:

$$L_{bce} = -\frac{1}{n}\sum_{i=1}^{n}(y_i \log p_i + (1 - y_i)\log(1 - p_i)) \tag{4}$$

Dice coefficient loss (DICE loss) is more sensitive to imbalanced datasets. Therefore, we introduce DICE loss as the other part of the loss function to weaken the imbalanced situation. The DICE loss is defined as:

$$L_{dice} = 1 - \frac{2 \times \left(\sum_{i=1}^{n} y_i \times p_i + smooth\right)}{\sum_{i=1}^{n} y_i + p_i + smooth} \tag{5}$$

In Equation (4) and (5), n represents the total pixel numbers, and y_i and $p_i$represent the values in the ground truth change map and the predicted change map, which are in a range of 0 to 1. In Equation (5), we added a parameter *smooth* and set its value as 1 to prevent the case of no change areas and zero-dominator situation in the DICE loss.

Therefore, the final loss function of the proposed EMS-CDNet is defined as follows:

$$L = L_{bce} + \lambda L_{dice} \tag{6}$$

where λ is used to control the weights of L_{bce}andL_{dice}. In the experiments, we set the value of λ as 0.5.

3. Experimental results and analysis

In this section, we analyze the experimental results to discuss the performance of the proposed EMS-CDNet. First, we depict the RSICD datasets used in Section 3.1, including two public datasets and our collected dataset. In addition, we introduce the evaluation metrics for quantitative measurement. Then, several SOTA RSICD methods are discussed, and their implementation details are illustrated. Our experimental results on the three datasets thereafter are displayed from the quantitative and qualitative perspectives.

3.1. Data description

To assess the performance of EMS-CDNet, we conducted experiments on two public RSICD datasets and our collected dataset. The two public datasets are a very-high-resolution remote sensing image dataset (the VHR dataset) (Lebedev et al. 2018) and the LEVIR-CD dataset (Chen and Shi 2020), where the former focuses on semantic change detection and the latter on building change detection.

The VHR dataset included 11 pairs of images collected from Google Earth and labeled by Lebedev et al. (2018). Seven pairs (with the size of 4725 $\times$ 2700 pixels) contained season-varying changes, and the other four (with the size of 1900 $\times$ 1000 pixels) contained manual creation changes. The resolution of these images varied from 0.03 m to 1.0 m, thus resulting in multi-scale changes. The images were randomly cropped into 16,000 patches with the size of 256 undefined pixels, of which 10,000 patches, 3,000 patches, and 3,000 patches were used for training, validation, and testing, respectively. Note that seasonal changes, such as grassland in summer and snow cover in winter, were not considered changes, making the dataset more challenging. Example images are shown in Figure 4.

The LEVIR-CD dataset (Chen and Shi 2020) collected 637 image pairs from Google Earth. These images were located in several cities, such as Austin and Lakeway in Texas, US, and their acquisition dates varied from 2002 to 2018. The images' resolution and size are 0.5 m and 1024 $\times$ 1024 pixels. Besides the seasonal changes, the images also suffered illumination changes, which made determining actual changes more challenging. The dataset was randomly divided into three parts for network training, where the ratio of the training, validation, and testing parts were 70%, 20%, and 10%, respectively. To avoid the over-fitting problem, we split each image pair into 16 patches with the size of 256 $\times$ 256 pixels. Several examples from the LEVIR-CD dataset are displayed in Figure 5.

The two public datasets were taken from Google Earth and were carefully cleaned up. Therefore, the influence of image registration errors was removed from the datasets. This is one of the reasons why most of the current models have performed well on these datasets but were unable to achieve good accuracy in practical applications. Therefore, we collected satellite image pairs to further observe EMS-CDNet's value in practical applications. The images in our collected dataset were acquired in 2017 with the world-view satellite and in 2018 with the GaoFen-2 (GF2) satellite. All the images were resampled to the resolution of 1 m. The newly built buildings, newly reclaimed paddies, and newly built highways were the main changes while some small objects, such as cars, were not considered as changes due to their relatively low resolution. We randomly cropped the images into patches with the size of 512 $\times$ 512 pixels and split them into

Figure 4. Example samples of the VHR dataset (Lebedev et al. 2018). In each line, from left to right, are image T_1, image T_2, and the labeled change map. The examples include the changes in buildings, roads, cars, and the not considered season-varying scenario.

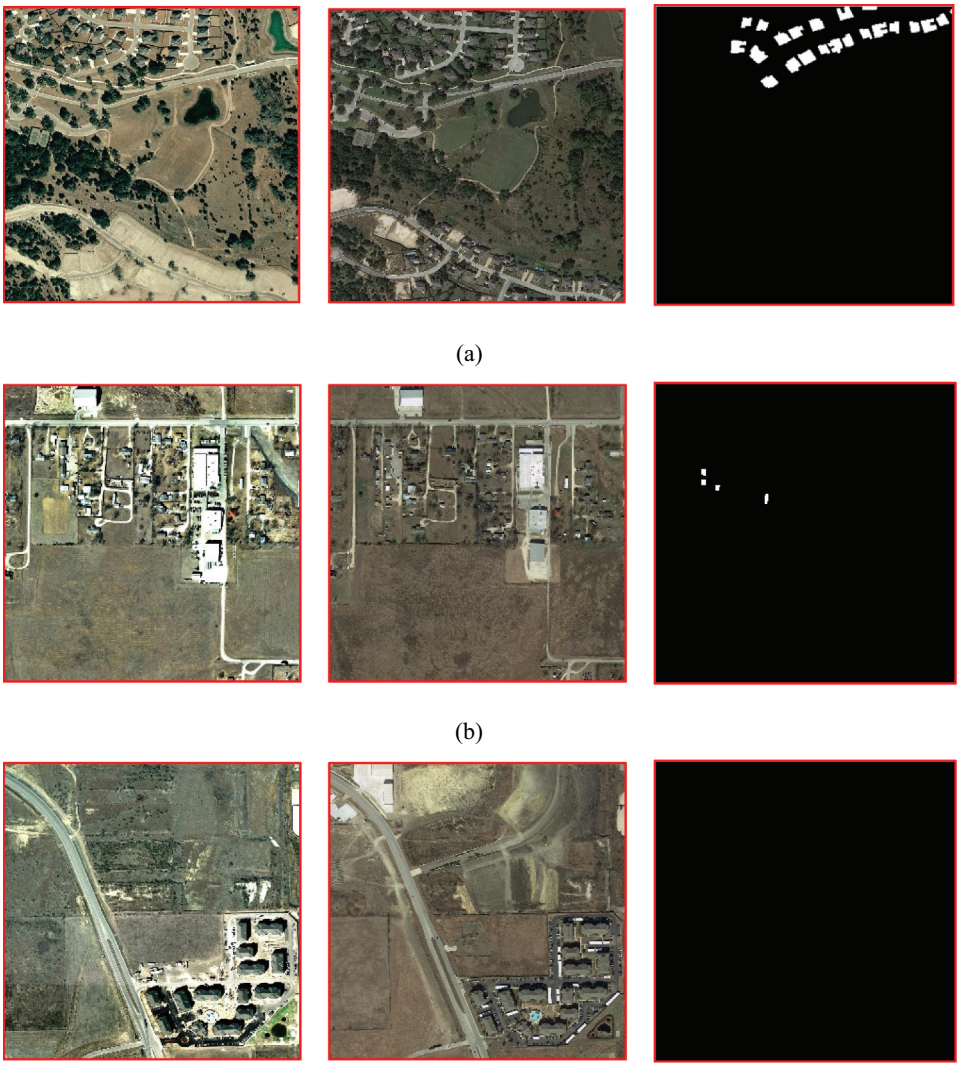


Figure 5. Example samples of the LEVIR-CD dataset (Chen and Shi 2020). In each line, from left to right, are image T_1, image T_2, and the labeled change map. The examples present the building update case, building decline case, and no change case, respectively.

training, validation, and testing parts with a ratio of 7:2:1. To ease the imbalanced distribution of the change patterns, the patches with no change areas were removed. Finally, our collected dataset included 3,542 patches, 1,013 patches, and 505 patches for training, validation, and testing. Some examples are shown in Figure 6.

To quantitatively measure the performance of EMS-CDNet, five evaluation metrics were selected to compare the difference between the predicted change maps and the labelled ground truth maps. The five metrics were precision (***P***), recall (***R***), f1-score (***F1***), overall accuracy (***OA***), and mean intersection over union (***mIoU***). ***P*** represents the ratio

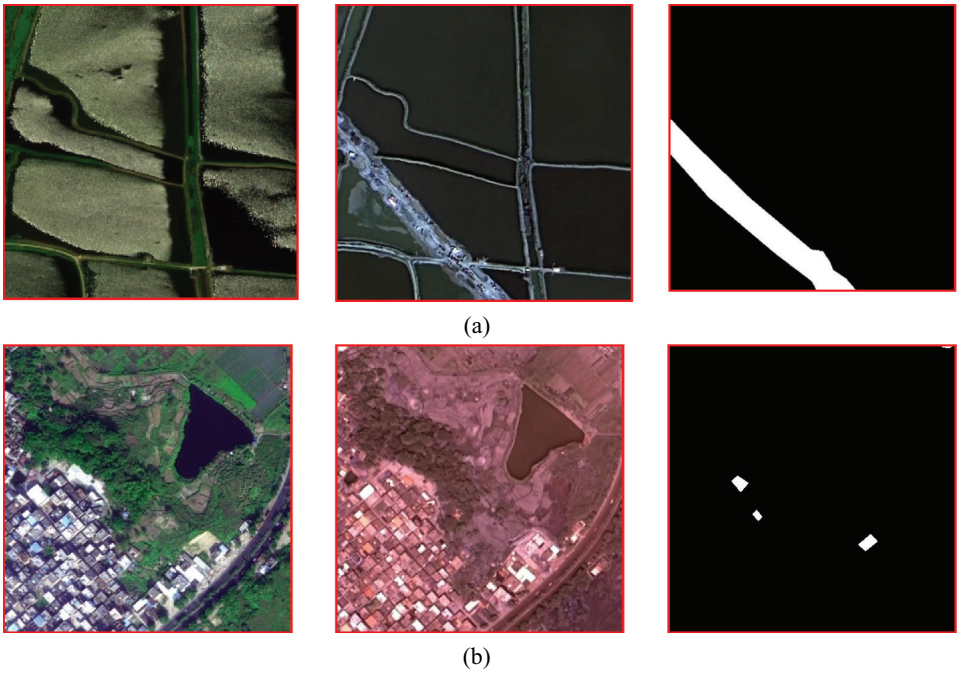


Figure 6. . example samples of the collected dataset. In each line, from left to right, are image T_1, image T_2, and the labeled change map. The examples present the changes in roads and buildings.

of the correctly predicted changed pixels to the whole predicted changed pixels, while ***R*** represents the ratio to the whole truly changed pixels. ***F1*** is the harmonic average of ***P*** and ***R***. ***OA*** indicates the proportion of the correctly predicted pixels to the whole pixels, and ***mIoU*** comprehensively considers the detection performance of the changed and unchanged areas. The calculation equations of the five metrics were as follows:

$$\mathrm{P} = \frac{TP}{TP + FP} \tag{7}$$

$$\mathrm{R} = \frac{TP}{TP + FN} \tag{8}$$

$$\mathrm{F1} = \frac{2 \times TP \times TN}{TP + FN} \tag{9}$$

$$\mathrm{OA} = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$\mathrm{mIoU} = \frac{TP}{FN + FP + TP} \tag{11}$$

where ***TP*** is the true positive value, ***TN*** is the true negative value, ***FP*** is the false positive value, and ***FN*** is the false negative value.

3.2. Comparison methods and implementation details

The following SOTA methods were selected to evaluate the performance of EMS-CDNet.

(1) FCN_EF, FC-Siam-conc, and FC-Siam-diff: Daudt et al. (2018) proposed three FCN-based networks for satellite RSICD tasks, which had different inputs in the encoder part. In our experiments, all three methods (FC_EF, FC-Siam-conc, and FC-Siam-diff) were selected for comparison. FC-EF stacked the image pairs as network input and utilized skip connections to complement the spatial details. Both FC-Siam-conc and the FC-Siam-diff were extensions of FC-EF, where they encoded the image features with different weight-shared strategies. FC-Siam-conc integrated the element-add features as encoding results while FC-Siam-diff took the element-minus features. These methods illustrated different feature aggregation approaches and were thus selected as the baseline in our experiments.
(2) FCN-PP: FCN-PP (Lei et al. 2019) is another comparison method we used. To overcome the drawbacks of global pooling strategies, FCN-PP embeds the pyramid pooling into the FCN backbone to enlarge the network's receptive field. It strengthened the context information and has proven to be effective for landslide detection. In our experiments, it was used to observe the effectiveness of the proposed MSF strategy.
(3) SNUNet-CD: SNUNet-CD (Fang et al. 2021) is a combination of the Siamese network and NestedUNet. Unlike most RSICD networks that usually focus on deep image features, SNUNet-CD adopted only the shallow-layer information to recover the sharp edges and to detect the changes in small targets. SNUNet-CD controlled the model's parameters by adjusting the width of the network. The parameters of SNUNet-CD/16 were in the same order of magnitude as the proposed EMS-CDNet. We selected it as a comparison to measure the effectiveness of the partition unit in our lightening models.
(4) Unet++_MSOF: Peng, Zhang, and Guan (2019) proposed an early-fusion RSICD network based on the Unet ++ architecture. For better illustration, we named it Unet++_MSOF in this paper. This work took advantage of the Unet++ backbone in multi-scale feature representation and refined the multiple network outputs with the multiple side-out fusion strategy (MSOF) and achieved high accuracy in many RSICD tasks. We selected it to measure the overall detection accuracy of EMS-CDNet.
(5) CLNet: CLNet (Zheng et al. 2021) is one of the newest Unet-series RSICD networks, in which a cross-layer block was designed to exploit multi-scale features and multi-level content information. Since sufficient image features were extracted and encoded, CLNet achieved superior detection performance and better accuracy-efficiency trade-off than the aforementioned methods. Thus, we further compared the differences between CLNet and EMS-CDNet to display the overall performance of EMS-CDNet.

All the methods were reproduced in the same experimental environment for a fair comparison. The experimental environment was an Ubuntu 18.04 workstation with a single NVIDIA GeForce GTX 1080Ti GPU with 12 G memory. All the methods were

reproduced under the Pytorch framework. In EMS-CDNet, the batch size was set as 20 for the VHR dataset and the LEVIR-CD dataset with the input size of {channels = 6, height = 256, width = 256} and 10 for the collected dataset with the input size of {channels = 6, height = 512, width = 512}. The initial learning rate was set at 0.001 and dropped by 10% when the training loss stopped decreasing for three epochs. Adam (Kingma, D. P., and J. Ba. 2014.) with default parameters ($\beta_1 = 0.9, \beta_2 = 0.999$) was selected as the optimizer, and the training procedure continued for 30 epochs in all the experiments.

3.3. Performance analysis of the VHR dataset

We first conducted experiments on the public VHR dataset to evaluate EMS-CDNet's performance in semantic change detection. Figures 7 and Figure 8 are visual comparisons of two typical testing areas that included changes in buildings, roads, and small targets. It can be seen that CLNet and EMS-CDNet obtained the best detection results as their predicted change maps were closely consistent with the labelled ground truth

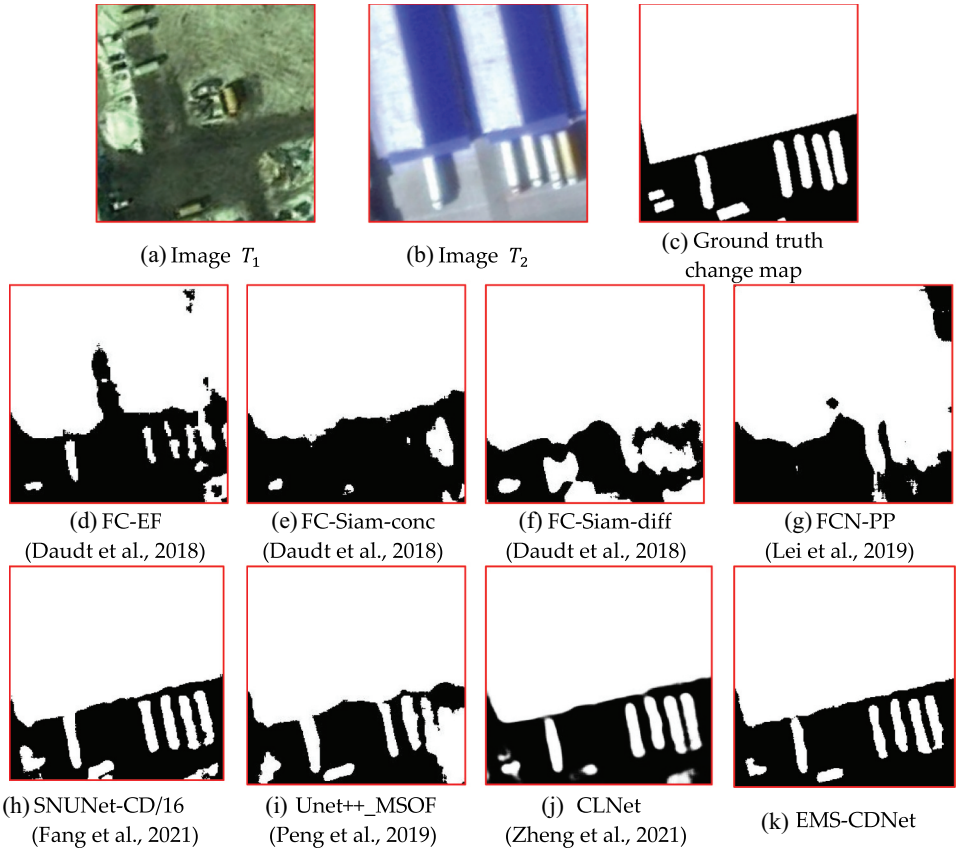


Figure 7. Visual comparisons for the first testing area in the VHR dataset. (a) image T_1; (b) image T_2; (c) ground truth change map; (d)-(k) change maps of FC-EF, FC-Siam-conc, FC-Siam-diff, FCN-PP, SNUNet-CD/16, Unet++_MSOF, CLNet, and the EMS-CDNet, where the changed pixels are labeled in white and the unchanged pixels are in black.

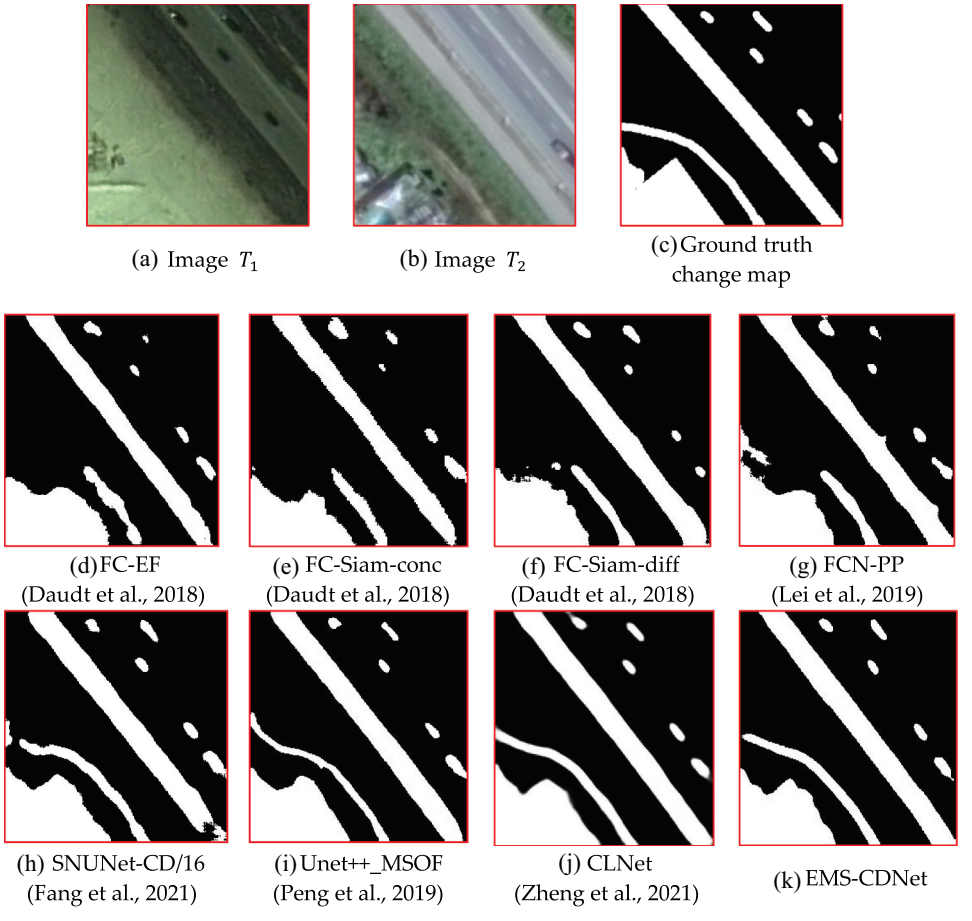


Figure 8. Visual comparisons for the second testing area in the VHR dataset. (a) image T_1; (b) image T_2; (c) ground truth change map; (d)-(k) change maps of FC-EF, FC-Siam-conc, FC-Siam-diff, FCN-PP, SNUNet-CD/16, Unet++_MSOF, CLNet, and the EMS-CDNet, where the changed pixels are labeled in white and the unchanged pixels are in black.

change map (see Figure 7(c), Figures 7(j), and Figure 7.). Figure 7 shows the changes in buildings where a new factory was built, along with several trucks parked nearby. In particular, EMS-CDNet and CLNet had fewer misdetection areas, and the detected building boundaries were more accurate than the other compared methods. The change maps' inner consistency and external inconsistency also were more obvious. The results of SNUNet-CD/16 and Unet++_MSOF were less accurate than EMS-CDNet (see the right corners in Figure 7(h) and Figure 7(i)). The remaining compared methods only detected rough change results and failed to recover the building boundaries. Figure 8 displays the change detection results for roads and small targets, where the changed objects were newly built highways and included moving cars. Note that while all the methods detected the changes, EMS-CDNet was able to draw a better change map. As for the narrow road in the left corner of this scenario, only Unet++_MSOF, CLNet, and EMS-CDNet detected the complete road. However, the other compared methods only detected a part of the road.

Table 1. Quantitative analysis of the VHR dataset (Best results are emphasized in bold).

Methods	Precision	Recall	F1	OA	mIoU
FC-EF (Daudt et al. 2018)	0.868	0.771	0.810	0.956	0.820
FC-Siam-conc (Daudt et al. 2018)	0.886	0.893	0.889	0.973	0.885
FC-Siam-diff (Daudt et al. 2018)	0.905	0.860	0.883	0.973	0.885
FCN-PP (Lei et al. 2019)	0.884	0.780	0.828	0.960	0.832
SNUNet-CD/16 (Fang et al. 2021)	0.921	0.900	0.911	0.967	0.902
Unet++_MSOF (Peng, Zhang, and Guan 2019)	0.920	0.910	0.910	0.978	0.907
CLNet (Zheng et al. 2021)	**0.947**	0.897	**0.921**	**0.981**	**0.921**
EMS-CDNet	0.923	**0.920**	**0.921**	0.980	0.918

As shown in Table 1, the evaluation metrics were calculated for quantitative comparison. Among the comparison methods, FC-EF obtained the worst quantitative results, which was caused by insufficient feature extraction, and it also was unable to detect detailed information in the testing areas (see Figure 7(d) and Figure 8(d)). When compared to FC-EF, FC-Siam-conc and FC-Siam-diff were able to exploit the feature correlations between the bi-temporal images, thereby achieving better qualitative results, with F1 values that increased by 1.8% and 7.3% and mIoU that increased by 1.2% and 6.5%, respectively. Benefiting from the extracted multi-scale features, FCN-PP overcame the drawbacks of global pooling and thus recovered more detailed information than FC-EF. SNUNet-CD/16, Unet++_MSOF, and CLNet further exploited the inner connection among different-level image features and integrated them into a unified framework. Therefore, their accuracy was further improved. For example, the F1 and mIoU of Unet++_MSOF increased by 2.2% and 2.1%, respectively, compared to FC-Siam-conc. In addition, their generated change maps were more complete than that of the aforementioned methods.

Note that CLNet achieved the best performance among all the methods. However, the performance of EMS-CDNet was similar to CLNet. For example, CLNet achieved the highest precision value (0.947), while EMS-CDNet achieved the highest recall value (0.920). Thus, their performances on F1 were essentially identical. As for the OA and mIoU, although CLNet obtained higher values than EMS-CDNet, the difference between CLNet and EMS-CDNet was very slight. Our experimental results demonstrated that EMS-CDNet achieved SOTA performance on the semantic change detection tasks and further indicated the effectiveness of our MSF strategy for complex feature representation.

3.4. Performance analysis of the LEVIR-CD dataset

We conducted experiments on the LEVIR-CD dataset to evaluate the EMS-CDNet's detection performance for buildings and small targets. Figures 9 and 10 depict the visual comparison of two typical testing areas. As shown in Figure 9, the change map produced by FC-Siam-conc was the worst among the comparison methods because the buildings were not detected completely (see Figure 9e). The other methods performed well on the testing samples and the difference among their generated change maps was very slight. The change maps of FCN-PP and Unet++_MSOF included more false detections than the other methods. Although the generated change maps of the other compared methods were similar, their detection results were not as complete as EMS-CDNet.

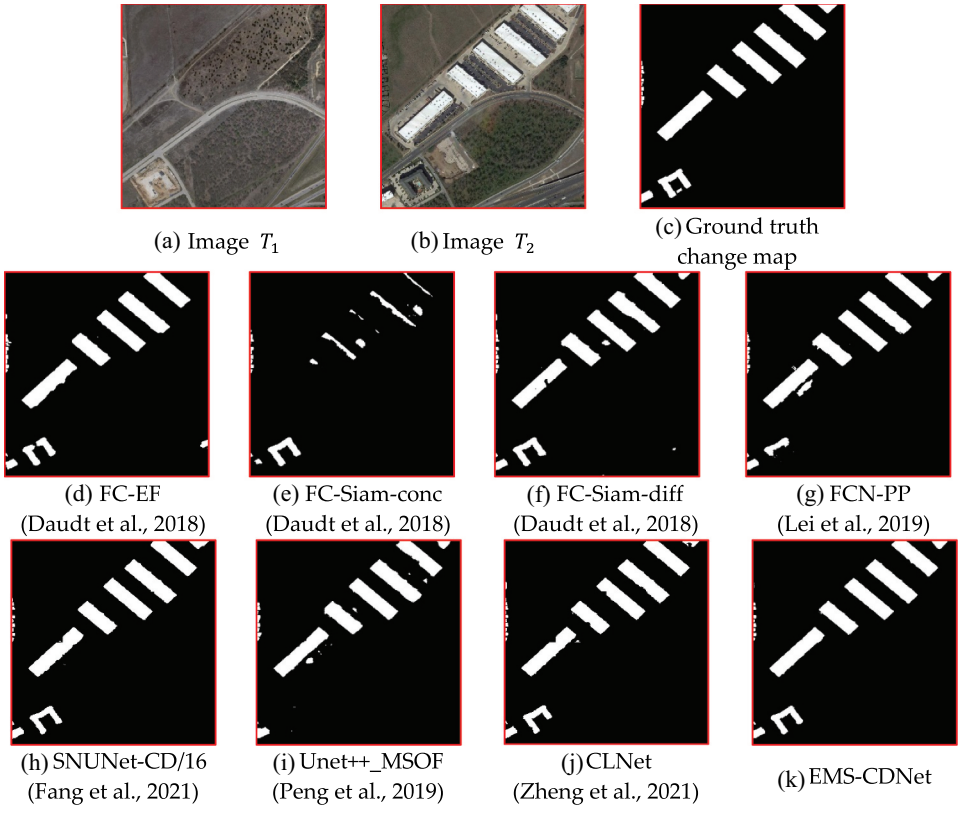


Figure 9. Visual comparisons for the LEVIR-CD dataset. (a) image T_1; (b) image T_2; (c) ground truth change map; (d)-(k) change maps of FC-EF, FC-Siam-conc, FC-Siam-diff, FCN-PP, SNUNet-CD/16, Unet++_MSOF, CLNet, and the EMS-CDNet, where the changed pixels are labeled in white and the unchanged pixels in black.

Table 2 lists the evaluation metrics we calculated for our quantitative comparisons. The quantitative results, as shown in Table 2, were in line with the visual performance, wherein FC-Siam-conc obtained the lowest F1 values. As shown in Table 2, EMS-CDNet achieved the highest precision, while its recall was lower than the other methods, except for FC-Siam-diff. Since the precision and recall were sensitive to the imbalanced data, we took the F1 as the indicator to compare the models' differences. Table 2 shows that CLNet achieved the best F1 followed by EMS-CDNet. Our experimental results demonstrated that EMS-CDNet could achieve SOTA performance on the building change detection tasks and further indicated its model's robustness.

3.5. Performance analysis of the collected dataset

The bi-temporal images for RSICD were captured with different sensors and from different perspectives. Therefore, the registration errors between the image pairs also were influenced by the detection accuracy of the models. However, both the VHR dataset and the LEVIR-CD dataset were collected from Google Earth and carefully cleaned up. The registration errors were artificially mitigated and therefore are one of the reasons why the

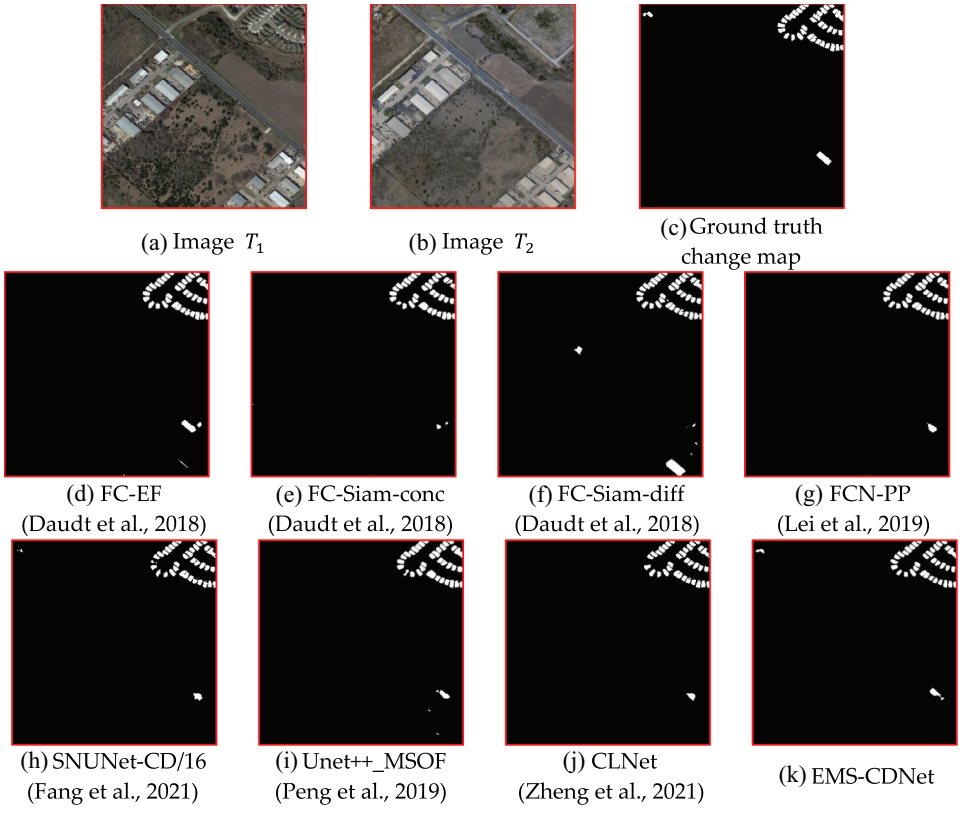


Figure 10. Visual comparisons for the LEVIR-CD dataset. (a) image T_1; (b) image T_2; (c) ground truth change map; (d)-(k) change maps of FC-EF, FC-Siam-conc, FC-Siam-diff, FCN-PP, SNUNet-CD/16, Unet++_MSOF, CLNet, and the EMS-CDNet, where the changed pixels are labeled in white and the unchanged pixels in black.

models trained on the public datasets did not perform well for actual RSICD tasks. Therefore, we collected satellite image pairs with registration errors and conducted another group of experiments to observe their performance under such situations.

Figures 11 and 12 display the changes in newly built buildings and widened roads, which were regarded as urban expansion monitors. Our visual comparison results indicated that EMS-CDNet achieved the best performance among all the methods and also generated the most complete change map and achieved a better view of the compact objects. The results of CLNet were the closest to EMS-CDNet, but CLNet introduced more false detections than EMS-CDNet (see the right corners of Figure 11j and Figure11k). In addition, EMS-CDNet also performed better in recovering narrow and small targets than CLNet; for example, EMS-CDNet recovered the buildings better in the bottom area of Figure 12. Unet++_MSOF obtained less accurate RSICD results compared to EMS-CDNet; for example, it failed to detect the changed buildings (see the right corner of Figure 11i). The other compared methods obtained worse results than EMS-CDNet as they could not detect the complete road changes in the test areas (see Figures 12(d) through Figure 12f).

The experimental results indicate that EMS-CDNet was more robust to image registration errors than the other compared methods mainly because of the application of our

Table 2. Quantitative analysis of the LEVIR-CD dataset (Best results are emphasized in bold).

Methods	Precision	Recall	F1	OA	mIoU
FC-EF (Daudt et al. 2018)	75.5	70.4	0.729	92.5	0.745
FC-Siam-conc (Daudt et al. 2018)	76.3	71.6	0.739	97.5	0.780
FC-Siam-diff (Daudt et al. 2018)	76.6	70.7	0.735	97.4	0.777
FCN-PP (Lei et al. 2019)	76.3	71.8	0.740	89.9	0.734
SNUNet-CD/16 (Fang et al. 2021)	85.8	86.5	0.862	96.7	0.860
Unet++_MSOF (Peng, Zhang, and Guan 2019)	86.7	86.9	0.868	98.5	0.821
CLNet (Zheng et al. 2021)	89.8	**90.3**	**0.900**	**98.9**	**0.886**
EMS-CDNet	**91.5**	85.0	0.880	97.4	0.863

MSF strategy. Unlike the other networks that only extracted features from the raw images, our MSF strategy enabled the extraction of multi-scale images for network input. In addition, the features extracted from the down-sampled branches were consistently the same size as the input images, thus avoiding the loss of structure details. Therefore, the

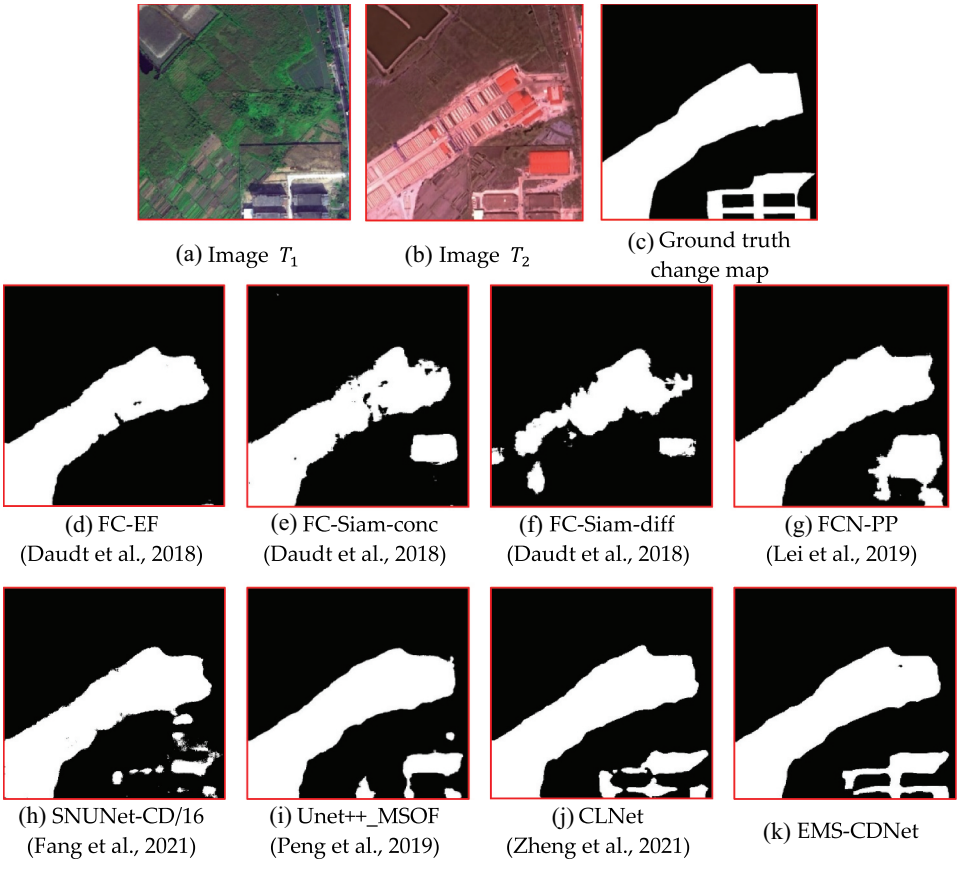


Figure 11. Visual comparisons for the collected dataset. (a) image T_1; (b) image T_2; (c) ground truth change map; (d)-(j) results of FC-EF, FC-Siam-conc, FC-Siam-diff, FCN-PP, SNUNet-CD/16, Unet++_MSOF, CLNet, and EMS-CDNet, where the changed pixels are labeled in white and the unchanged pixels in black.

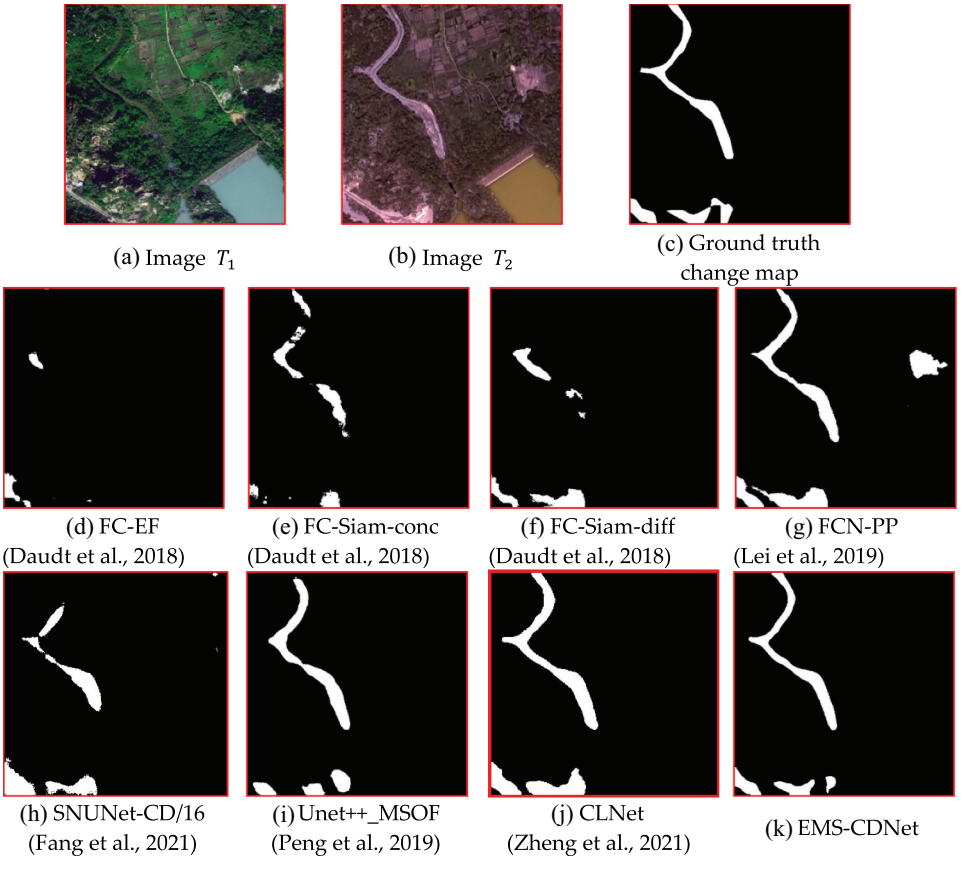


Figure 12. Visual comparisons for the collected dataset. (a) image T_1; (b) image T_2; (c) ground truth change map; (d)-(j) results of FC-EF, FC-Siam-conc, FC-Siam-diff, FCN-PP, SNUNet-CD/16, Unet++_MSOF, CLNet, and EMS-CDNet, where the changed pixels are labeled in white and the unchanged pixels are in black.

side-effect of image registration errors was alleviated by the multi-scale input images, and thus the network achieved better overall performance.

Our quantitative assessment of our collected dataset is shown in Table 3. Compared to the public VHR dataset and LEVIR-CD dataset, the image pairs in our collected dataset contained image registration errors. Therefore, the quantitative results of all the compared methods significantly decreased compared to their performance on the public datasets. EMS-CDNet and CLNet still outperformed the other compared methods, demonstrating their robustness. In addition, the other method tended to misclassify the changed pixels to unchanged pixels and thus obtained much higher precision values than recall values, as did FC-Siam-diff and FCN-PP. However, this situation did not occur in the results of EMS-CDNet. In addition, EMS-CDNet achieved the best performance on four metrics, especially the mIoU metric, which further verified its robustness to the image registration errors.

Table 3. Quantitative analysis of the collected dataset (best results are emphasized in bold).

Methods	Precision	Recall	F1	OA	mIoU
FC-EF (Daudt et al. 2018)	77.0	42.0	0.550	97.0	0.670
FC-Siam-conc (Daudt et al. 2018)	68.0	40.0	0.510	96.0	0.650
FC-Siam-diff (Daudt et al. 2018)	57.0	41.0	0.480	96.0	0.640
FCN-PP (Lei et al. 2019)	59.0	45.0	0.510	96.0	0.650
SNUNet-CD/16 (Fang et al. 2021)	49.4	58.2	0.534	96.1	0.662
Unet++_MSOF (Peng, Zhang, and Guan 2019)	65.0	69.0	0.670	96.0	0.730
CLNet (Zheng et al. 2021)	**79.0**	69.4	0.734	**97.7**	0.793
EMS-CDNet	**79.0**	**70.0**	**0.740**	97.0	**0.806**

4. Discussion

In this section, we first present our experiments to investigate the effectiveness of our partition unit for lightening the model. Then, we evaluate the model's performance based on the accuracy-efficiency trade-off. Among the selected comparison methods, CLNet achieved the best accuracy performance in most cases. We further compare EMS-CDNet to CLNet, from the viewpoint of decreased accuracy to improved efficiency in order to determine its all-around performance for RSICD tasks.

4.1. Effectiveness evaluation of the partition unit

We conducted experiments on the VHR dataset to illustrate the effectiveness of our partition unit. We used the Unet structure with the proposed MSF strategy as the baseline. Several commonly used lightening strategies were selected as comparisons to evaluate the effectiveness of the partition unit, such as group convolution (Cohen and Welling 2016) and separate convolution (Chollet 2017). We calculated the model parameters and the whole processing time for 3,000 testing image pairs. We selected the mIoU to measure model accuracy. As shown in Table 4, the parameters, the mIoU, and the whole processing time of the baseline model were 5.58 M, 0.866, and 2430s, respectively.

We first applied group convolution and separable convolution. As shown in Table 4, using group convolution and separable convolution significantly reduced the model parameters. Compared to the baseline model, the parameters were reduced by 46.4% and 46.7%, respectively. Although the whole processing time was cut in half, the accuracy of the two models declined as well. We integrated the channel shuffle strategy (Zhang et al. 2018) and separated convolution, and then embedded the combination module into the Unet. Since the channel shuffle strategy exploits the channel correlation information, this model achieved improvement in accuracy compared to the models that only apply separable convolution. In addition, the whole processing time was further decreased to 958s.

After embedding the partition unit into the baseline model, the mIoU of the model increased to 0.918, which was the opposite of the other models. The main reason for this increase was that the design of the partition unit maintained the channel-wise contextual information and facilitated the fusion of the multi-scale features. The whole processing time of the proposed model also decreased to half of the baseline model. As shown in

Table 4, the processing time of each model was positively correlated to the model's parameters. Therefore, the processing time of the proposed model was slightly longer than the other models. In our opinion, the extra processing time was acceptable and worthwhile because of the significant accuracy improvement. For example, the mIoU of the proposed model was 6.9% higher than the model using channel shuffle strategy and separable convolution. At the same time, the extra processing time was only extended by 82s for 3,000 image pairs.

4.2. Performance on accuracy-efficiency trade-off

We conducted additional experiments to further observe EMS-CDNet's performance on the accuracy-efficiency trade-off. Then, we counted the model parameters, the mIoU, and the whole processing time of all the comparison methods. We classified the comparisons into two categories according to each model's parameters: 1) lightweight models (FC-EF, FC-Siam-conc, FC-Siam-diff, and SNUNet-CD/16) and 2) heavyweight models (FCN-PP, Unet++_MSOF, and CLNet).

As shown in Table 5, the detection accuracy of the heavyweight models was much better than that of the lightweight models (except for the abnormal FCN-PP). For example, the mIoU of Unet++_MSOF was 0.730 on our collected dataset, while the value for FC_EF was only 0.670. Although the heavyweight models achieved better detection accuracies, their processing time was much longer than the lightweight models. For example, the processing time of CLNet was 1550s on the VHR dataset, while the SNUNet-CD/16's processing time was only 972s. From the experimental results, we concluded that the heavyweight models achieved higher accuracy by sacrificing their efficiency, but the situation was just the opposite for the lightweight models because the heavyweight models always design complex feature extractors for sufficient information interpretation. In contrast, lightweight models focus on processing time and thus experience accuracy problems.

As shown in Table 5, EMS-CDNet achieved identical accuracy to the heavyweight models while its processing time was approximately equal to the lightweight models. For example, its mIoU and processing time were 0.806 and 295s on our collected dataset. The highest mIoU among the heavyweight models was 0.793, which was achieved by CLNet. However, the processing time of CLNet was 504s. The shortest processing time of the lightweight models was 237s, while the mIoU value of FC_EF was only 0.670. Therefore, we concluded that EMS-CDNet achieved a better accuracy-efficiency trade-off than the comparison methods.

Table 4. Effectiveness evaluation of the proposed partition unit.

Baseline	Model Lightening Operations	Para.	mIoU	Time
Unet	None	5.58M	0.866	2430s
+MSF	Group convolution (Cohen and Welling 2016)	3.55M	0.847	1024s
	Separable convolution (Chollet 2017)	3.53M	0.849	1029s
	Separable convolution (Chollet 2017) & channel shuffle (Zhang et al. 2018)	3.30M	0.854	958s
	Partition unit	3.88M	0.918	1040s

Table 5. . effectiveness and efficiency comparison.

	Methods	Para.	VHR dataset		LEVIR-CD dataset		collected dataset	
			mIoU	Time	mIoU	Time	mIoU	Time
Lightweight models	FC-EF (Daudt et al. 2018)	1.44M	0.820	927s	0.745	316s	0.670	237s
	FC-Siam-conc (Daudt et al. 2018)	1.63M	0.885	1093s	0.780	374s	0.650	283s
	FC-Siam-diff (Daudt et al. 2018)	1.44M	0.885	1021s	0.777	366s	0.640	275s
	SNUNet-CD/16 (Fang et al. 2021)	3.01M	0.902	972s	0.734	348s	0.662	252s
Heavyweight Models	FCN-PP (Lei et al. 2019)	9.94M	0.832	1450s	0.860	423s	0.650	397s
	Unet++_MSOF (Peng, Zhang, and Guan 2019)	9.70M	0.907	5570s	0.821	1131s	0.730	938s
	CLNet (Zheng et al. 2021)	8.00M	0.921	1550s	**0.886**	688s	0.793	504s
Our Model	EMS-CDNet	3.88M	0.918	1040s	0.863	380s	0.806	295s

4.3. Performance compared to CLNet

As confirmed in the literature about CLNet(Zheng et al. 2021), CLNet achieved a better accuracy-efficiency trade-off than the comparison methods. Therefore, we further compared EMS-CDNet to CLNet both on the public VHR dataset and the collected dataset to observe the efficiency improvement and the accuracy decrease. As shown in Table 6, there was a slight accuracy difference between EMS-CDNet and CLNet. For example, the F1, OA, and mIoU of EMS-CDNet only decreased by 0%, 0.1%, and 0.3% to CLNet in the VHR dataset. On the collected dataset, the OA of EMS-CDNet decreased by 0.7% to CLNet. However, the F1 and mIoU of EMS-CDNet increased by 0.6% and 1.3%, thereby benefitting from EMS-CDNet's advantages in suppressing image registration errors. In particular, the whole processing time of EMS-CDNet was shortened by 33%, and 41.5% for the three datasets, which significantly improved EMS-CDNet's values in practical and emergency RSICD tasks. Therefore, we believe the slight accuracy decrease was acceptable under the significantly shortened processing time.

(↑indicates improvement and ↓indicates decrease)

4.4. Limitations of the EMS-CDNet

Although the above three subsections comprehensively illustrated the superiority of EMS-CDNet, this work experienced the following limitations:

Table 6. Performance difference compared to the CLNet.

	VHR dataset				collected dataset			
		Accuracy				Accuracy		
	Time	F1	OA	mIoU	Time	F1	OA	mIoU
CLNet (Zheng et al. 2021)	1550s	0.921	0.981	0.921	504s	0.734	0.977	0.793
EMS-CDNet	1040s	0.921	0.980	0.918	295s	0.740	0.970	0.806
Rate	**↑33%**	**↓0%**	**↓0.1%**	**↓0.3%**	**↑41.5%**	**↑0.6%**	**↓0.7%**	**↑1.3%**

(1) The boundaries of the buildings obtained by EMS-CDNet were not sharp enough. We believe the down-sampling and pooling operations we used inevitably lost the structured information. Therefore, we intend to investigate boundary refinement strategies to address the issue and introduce them to refine the predicted change maps.
(2) The requirements of numerous labelled training samples limited the application of EMS-CDNet. Therefore, recently developed techniques, such as weakly supervised and self-supervised learning, will be investigated in the future.
(3) Our EMS-CDNet focuses on binary RSICD tasks rather than semantic ones, which is not applicable for some particular RSICD tasks.

5. Conclusions

In this paper, we introduced a novel method for RSICD tasks, called EMS-CDNet, which includes two novel elements (an MSF strategy and a partition unit) that were shown to improve the model's performance in practical applications. Our MSF strategy enables sufficient feature extraction and its multi-scale image inputs make it more robust to image registration errors than do single-scale inputs. Our partition unit significantly lightened our EMS-CDNet model while introducing a negligible decrease in accuracy. Thus, it achieved a better accuracy-efficiency trade-off than the comparison methods. For example, while the accuracy results for CLNet and EMS-CDNet were identical, EMS-CDNet experienced 33% efficiency gains over CLNet.

Our experimental results show that EMS-CDNet achieved excellent results on two public datasets and our collected dataset of satellite images, with an F1 of 92.1%, 86.3%, and 74% on the VHR dataset, the LEVIR-CD dataset, and the collected dataset. Additional experimental results also confirmed the value of applying our EMS-CDNet in practical applications. In our future work, we will continue investigating EMS-CDNet's performance on other challenging images and refining the network structure. We also will exploit the feasibility of integrating the proposed network into a multi-task framework to expand its application values.

Acknowledgments

This research was funded in part by the National Natural Science Foundation of China under Grants No. 42030102, 41801386, and 42001406 ; the Fund for Innovative Research Groups of the Hubei Natural Science Foundation under Grant No. 2020CFA003; Major special projects of Guizhou [2022] 001; and the China Postdoctoral Science Foundation under Grant No. 2020M672416. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University. The authors are grateful to the creators of the change detection datasets used in this research for providing them as public datasets. The peer reviewers' comments and suggestions about this manuscript also were greatly appreciated.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by National Natural Science Foundation of China [42030102, 41801386,42001406]; Innovative Research Groups of the Hubei Natural Science Foundation [2020CFA003]; Major special projects of Guizhou [[2022]001], the China Postdoctoral Science Foundation [2020M672416].

Data availability statement

The VHR dataset and the LEVIR-CD dataset that support the findings of this work are available at https://drive.google.com/file/d/1GX656JqqOyBi_Ef0w65kDGVto-nHrNs9 and https://justchenhao.github.io/LEVIR/, which were produced by Lebedev et al. (2018) and Chen and Shi (2020), respectively. The collected dataset that supports the findings of this article is not available because the dataset producer declined to make the data public due to legal restrictions.

References

Bock, M., P. Xofis, J. Mitchley, G. Rossner, and M. Wissen. 2005. "Object-Oriented Methods for Habitat Mapping at Multiple Scales–case Studies from Northern Germany and Wye Downs, UK." *Journal for Nature Conservation* 13 (2–3): 75–89. doi:10.1016/j.jnc.2004.12.002.

Celik, T. 2009. "Unsupervised Change Detection in Satellite Images Using Principal Component Analysis and *K-Means* Clustering." *IEEE Geoscience and Remote Sensing Letters* 6 (4): 772–776.

Chen, G., G. J. Hay, L. M. Carvalho, and M. A. Wulder. 2012. "Object-Based Change Detection." *International Journal of Remote Sensing* 33 (14): 4434–4457.

Chen, H., and Z. Shi. 2020. "A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection." *Remote Sensing* 12 (10): 1662. doi:10.3390/rs12101662.

Chollet, F. (2017). "Xception: Deep Learning with Depthwise Separable Convolutions." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Honolulu, HI, USA. 1251–1258.

Cohen, T., and M. Welling (2016, June). "Group Equivariant Convolutional Networks." In *International conference on machine learning*. New York City, United States. 2990–2999. PMLR.

Dalla Mura, M., J. A. Benediktsson, F. Bovolo, and L. Bruzzone. 2008. "An Unsupervised Technique Based on Morphological Filters for Change Detection in Very High Resolution Images." *IEEE Geoscience and Remote Sensing Letters* 5 (3): 433–437.

Daudt, R. C., B. Le Saux, A. Boulch, and Y. Gousseau. 2018. "High Resolution Semantic Change Detection." *arXiv preprint arXiv:1810.08452*

Erener, A., and H. S. Düzgün. 2009. "A Methodology for Land Use Change Detection of High Resolution Pan Images Based on Texture Analysis." *Italian Journal of Remote Sensing* 41 (2): 47–59. doi:10.5721/ItJRS20094124.

Fang, S., K. Li, J. Shao, and Z. Li. 2021. "SNUNet-CD: A Densely Connected Siamese Network for Change Detection of VHR Images." *IEEE Geoscience and Remote Sensing Letters* 19: 1–5.

Gamanya, R., P. De Maeyer, and M. De Dapper. 2009. "Object-Oriented Change Detection for the City of Harare, Zimbabwe." *Expert Systems with Applications* 36 (1): 571–588.

Girshick, R. (2015). "Fast R-Cnn." In *Proceedings of the IEEE international conference on computer vision*, Santiago, Chile. 1440–1448.

Gong, M., Y. Yang, T. Zhan, X. Niu, and S. Li. 2019. "A Generative Discriminatory Classified Network for Change Detection in Multispectral Imagery." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12 (1): 321–333. doi:10.1109/JSTARS.2018.2887108.

Gong, M., J. Zhao, J. Liu, Q. Miao, and L. Jiao. 2015. "Change Detection in Synthetic Aperture Radar Images Based on Deep Neural Networks." *IEEE Transactions on Neural Networks and Learning Systems* 27 (1): 125–138.

Howard, A. G., M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, Marco Andreetto, Adam, H. 2017. "Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications." *arXiv preprint arXiv:1704.04861*

Huang, W., S. Zhang, and H. H. Wang (2020, May). "Efficient GAN-Based Remote Sensing Image Change Detection Under Noise Conditions." In *International conference on image processing and capsule networks* (pp. 1–8). Springer, Cham.

Hussain, M., D. Chen, A. Cheng, H. Wei, and D. Stanley. 2013. "Change Detection from Remotely Sensed Images: From Pixel-Based to Object-Based Approaches." *ISPRS Journal of Photogrammetry and Remote Sensing* 80: 91–106. doi:10.1016/j.isprsjprs.2013.03.006.

Jin, S., L. Yang, P. Danielson, C. Homer, J. Fry, and G. Xian. 2013. "A Comprehensive Change Detection Method for Updating the National Land Cover Database to Circa 2011." *Remote Sensing of Environment* 132: 159–175. doi:10.1016/j.rse.2013.01.012.

Ji, S., Y. Shen, M. Lu, and Y. Zhang. 2019. "Building Instance Change Detection from Large-Scale Aerial Images Using Convolutional Neural Networks and Simulated Samples." *Remote Sensing* 11 (11): 1343. doi:10.3390/rs11111343.

Kingma, D. P., and J. Ba. 2014. "Adam: A Method for Stochastic Optimization." *arXiv preprint arXiv:1412.6980*.

Lebedev, M. A., Y. V. Vizilter, O. V. Vygolov, V. A. Knyaz, and A. Y. Rubis. 2018. "Change Detection in Remote Sensing Images Using Conditional Adversarial Network." *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences* 42 (2): 565–571.

Lefebvre, A., T. Corpetti, and L. Hubert-Moy (2008, July). Object-Oriented Approach and Texture Analysis for Change Detection in Very High Resolution Images. In *IGARSS 2008-2008 IEEE International Geoscience and Remote Sensing Symposium* (Vol. 4, pp. IV–663). IEEE.

Lei, T., Y. Zhang, Z. Lv, S. Li, S. Liu, and A. K. Nandi. 2019. "Landslide Inventory Mapping from Bitemporal Images Using Deep Convolutional Neural Networks." *IEEE Geoscience and Remote Sensing Letters* 16 (6): 982–986. doi:10.1109/LGRS.2018.2889307.

Lin, T. Y., P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie (2017). "Feature Pyramid Networks for Object Detection." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Honolulu, HI, USA. 2117–2125.

Long, J., E. Shelhamer, and T. Darrell (2015). "Fully Convolutional Networks for Semantic Segmentation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Boston, MA, USA. 3431–3440.

Lu, D., S. Hetrick, E. Moran, and G. Li. 2010. "Detection of Urban Expansion in an Urban-Rural Landscape with Multitemporal QuickBird Images." *Journal of Applied Remote Sensing* 4 (1): 041880.

Lv, Z., H. Huang, L. Gao, J. A. Benediktsson, M. Zhao, and C. Shi. 2022c. "Simple Multiscale UNet for Change Detection with Heterogeneous Remote Sensing Images." *IEEE Geoscience and Remote Sensing Letters* 19:1558-0571. doi:10.1109/LGRS.2022.3173300.

Lv, Z., G. Li, Z. Jin, J. A. Benediktsson, and G. M. Foody. 2020. "Iterative Training Sample Expansion to Increase and Balance the Accuracy of Land Classification from VHR Imagery." *IEEE Transactions on Geoscience and Remote Sensing* 59 (1): 139–150.

Lv, Z., F. Wang, G. Cui, J. A. Benediktsson, T. Lei, and W. Sun. 2022b. "Spatial–spectral Attention Network Guided with Change Magnitude Image for Land Cover Change Detection Using Remote Sensing Images." *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–12.

Lv, Z., F. Wang, W. Sun, Z. You, N. Falco, and J. A. Benediktsson (2022a). "Landslide Inventory Mapping on VHR Images via Adaptive Region Shape Similarity." IEEE Transactions on Geoscience and Remote Sensing 60:1558-0644. doi:10.1109/TGRS.2022.3204834.

Lv, P., Y. Zhong, J. Zhao, H. Jiao, and L. Zhang. 2016. "Change Detection Based on a Multifeature Probabilistic Ensemble Conditional Random Field Model for High Spatial Resolution Remote Sensing Imagery." *IEEE Geoscience and Remote Sensing Letters* 13 (12): 1965–1969. doi:10.1109/LGRS.2016.2619163.

Lyu, H., H. Lu, and L. Mou. 2016. "Learning a Transferable Change Rule from a Recurrent Neural Network for Land Cover Change Detection." *Remote Sensing* 8 (6): 506. doi:10.3390/rs8060506.

Mehta, S., M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi (2018). "Espnet: Efficient Spatial Pyramid of Dilated Convolutions for Semantic Segmentation." In *Proceedings of the european conference on computer vision (ECCV)* Munich, Germany. 552–568.

Mou, L., L. Bruzzone, and X. X. Zhu. 2018. "Learning Spectral-Spatial-Temporal Features via a Recurrent Convolutional Neural Network for Change Detection in Multispectral Imagery." *IEEE Transactions on Geoscience and Remote Sensing* 57 (2): 924–935.

Nielsen, A. A. 2007. "The Regularized Iteratively Reweighted MAD Method for Change Detection in Multi-And Hyperspectral Data." *IEEE Transactions on Image Processing* 16 (2): 463–478.

Peng, D., Y. Zhang, and H. Guan. 2019. "End-To-End Change Detection for High Resolution Satellite Images Using Improved UNet++." *Remote Sensing* 11 (11): 1382. doi:10.3390/rs11111382.

Ronneberger, O., P. Fischer, and T. Brox (2015, October). U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234–241). Springer, Cham.

Saha, S., F. Bovolo, and L. Bruzzone. 2019. "Unsupervised Deep Change Vector Analysis for Multiple-Change Detection in VHR Images." *IEEE Transactions on Geoscience and Remote Sensing* 57 (6): 3677–3693.

Sandler, M., A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen (2018). "Mobilenetv2: Inverted Residuals and Linear Bottlenecks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Salt Lake City, UT, USA. 4510–4520.

Song, A., J. Choi, Y. Han, and Y. Kim. 2018. "Change Detection in Hyperspectral Images Using Recurrent 3D Fully Convolutional Networks." *Remote Sensing* 10 (11): 1827. doi:10.3390/rs10111827.

Tomowski, D., M. Ehlers, and S. Klonus (2011, April). "Colour and Texture Based Change Detection for Urban Disaster Analysis." In *2011 Joint Urban Remote Sensing Event*. Munich, Germany. 329–332. IEEE.

Wang, F., and Y. J. Xu. 2010. "Comparison of Remote Sensing Change Detection Techniques for Assessing Hurricane Damage to Forests." *Environmental Monitoring and Assessment* 162 (1): 311–326.

Wang, Q., Z. Yuan, Q. Du, and X. Li. 2018. "GETNET: A General End-To-End 2-D CNN Framework for Hyperspectral Image Change Detection." *IEEE Transactions on Geoscience and Remote Sensing* 57 (1): 3–13.

Wen, D., X. Huang, F. Bovolo, J. Li, X. Ke, A. Zhang, and J. A. Benediktsson. 2021. "Change Detection from Very-High-Spatial-Resolution Optical Remote Sensing Images: Methods, Applications, and Future Directions." *IEEE Geoscience and Remote Sensing Magazine* 9 (4): 68–101.

Wu, C., B. Du, X. Cui, and L. Zhang. 2017. "A Post-Classification Change Detection Method Based on Iterative Slow Feature Analysis and Bayesian Soft Fusion." *Remote Sensing of Environment* 199: 241–255. doi:10.1016/j.rse.2017.07.009.

Yang, L., G. Xian, J. M. Klaver, and B. Deal. 2003. "Urban Land-Cover Change Detection Through Sub-Pixel Imperviousness Mapping Using Remotely Sensed Data." *Photogrammetric Engineering & Remote Sensing* 69 (9): 1003–1010.

Zhang, Y., D. Peng, and X. Huang. 2017. "Object-Based Change Detection for VHR Images Based on Multiscale Uncertainty Analysis." *IEEE Geoscience and Remote Sensing Letters* 15 (1): 13–17.

Zhang, C., P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, and G. Liu. 2020. "A Deeply Supervised Image Fusion Network for Change Detection in High Resolution Bi-Temporal Remote Sensing Images." *ISPRS Journal of Photogrammetry and Remote Sensing* 166: 183–200.

Zhang, X., X. Zhou, M. Lin, and J. Sun (2018). "Shufflenet: An Extremely Efficient Convolutional Neural Network for Mobile Devices." In *Proceedings of the IEEE conference on computer vision and pattern recognition* Salt Lake City, UT, USA (pp. 6848–6856).

Zheng, Z., Y. Wan, Y. Zhang, S. Xiang, D. Peng, and B. Zhang. 2021. "CLNet: Cross-Layer Convolutional Neural Network for Change Detection in Optical Remote Sensing Imagery." *ISPRS Journal of Photogrammetry and Remote Sensing* 175: 247–267. doi:10.1016/j.isprsjprs.2021.03.005.

Zhou, Z., M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang. 2018. "Unet++: A Nested U-Net Architecture for Medical Image Segmentation." Stoyanov, Danail, Carneiro, Gustavo, Taylor, Zeike,

Syeda-Mahmood, Tanveer. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 3–11. Cham: Springer.

Zhou, W., A. Troy, and M. Grove. 2008. "Object-Based Land Cover Classification and Change Analysis in the Baltimore Metropolitan Area Using Multitemporal High Resolution Remote Sensing Data." *Sensors* 8 (3): 1613–1636. doi:10.3390/s8031613.

Zhu, X. X., D. Tuia, L. Mou, G. S. Xia, L. Zhang, F. Xu, and F. Fraundorfer. 2017. "Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources." *IEEE Geoscience and Remote Sensing Magazine* 5 (4): 8–36.