SNUNet-CD: A Densely Connected Siamese Network for Change Detection of VHR Images

Sheng Fang, Kaiyu Li, Jinyuan Shao, and Zhe Li

***Abstract*— Change detection is an important task in remote sensing (RS) image analysis. It is widely used in natural disaster monitoring and assessment, land resource planning, and other fields. As a pixel-to-pixel prediction task, change detection is sensitive about the utilization of the original position information. Recent change detection methods always focus on the extraction of deep change semantic feature, but ignore the importance of shallow-layer information containing high-resolution and fine-grained features, this often leads to the uncertainty of the pixels at the edge of the changed target and the determination miss of small targets. In this letter, we propose a densely connected siamese network for change detection, namely SNUNet-CD (the combination of Siamese network and NestedUNet). SNUNet-CD alleviates the loss of localization information in the deep layers of neural network through compact information transmission between encoder and decoder, and between decoder and decoder. In addition, Ensemble Channel Attention Module (ECAM) is proposed for deep supervision. Through ECAM, the most representative features of different semantic levels can be refined and used for the final classification. Experimental results show that our method improves greatly on many evaluation criteria and has a better tradeoff between accuracy and calculation amount than other state-of-the-art (SOTA) change detection methods.**

***Index Terms*— Change detection, deep learning, fully convolutional siamese network, remote sensing (RS) images.**

I. Introduction

THE objective of change detection is to detect pixels with "semantic change" between multitemporal remote sensing (RS) images acquired at different times and in the same area. There are many factors that may cause "semantic change," such as the deformation, relative motion, appearance, or disappearance of the object. The difficulty of change detection task is that the final change map should not contain "non-semantic change," such as changes caused by camera motion, sensor noise, or light change. Another difficulty in change detection is that the definition of "change" may vary depending on the application and the subjective consciousness of the person. For example, in many cases, bi-temporal images are obtained from different seasons, as shown in Fig. 1. "Change" is defined as changes in man-made facilities such as buildings and cars, while seasonal changes are regarded as interference factors. Therefore, many traditional change detection methods, such as Image Difference, change vector analysis (CVA) [2] and PCA & Kmeans [3], etc., which can achieve effective results in some simple scenarios, often perform poorly in these complex scenarios.

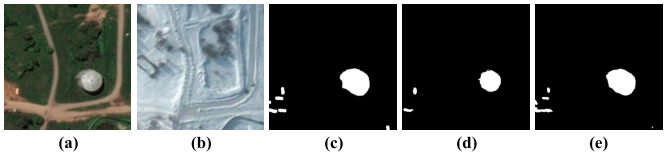

Fig. 1. Bi-temporal RS images with season-varying in CDD data set [1]. (a) and (b) are the original bitemporal images, (c) is ground truth, (d) is the result of FC-Siam-diff [5], a common method in change detection, and (e) is our result.

In recent years, many techniques and components of neural networks for scene segmentation have been used in change detection task to extract deeper representation. First, U-Net [4] takes the lead and establish a benchmark model; then, the siamese network is used and become the standard method for change detection [5]–[11]. To improve the performance of change detection, there are a lot of efforts on deep feature extraction and refinement. In [10], the pyramid model is used for extracting multiscale features; in [9] and [12], deep supervision is used to enhance the representation and discrimination capabilities of shallow features; and attention mechanisms are used to refine features and obtain better feature representations, such as spatial and channel attention in [9], self-attention in [10], and dual attention in [11], and so on.

Although these methods have achieved practical success, a common problem is that successive down-samplings cause the loss of accurate spatial position information, which often leads to the uncertainty of the pixels at the edge of the changed target and the determination miss of small targets, as shown in Fig. 1(d). Many studies [4], [13]–[15] indicate that shallow layers of the neural network contain fine-grained localization information, while deep layers contain coarse-grained semantic information. Inspired by DenseNet [16] and NestedUNet [14], [15], we design a densely

Manuscript received August 6, 2020; revised December 18, 2020; accepted January 29, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61502278, in part by the National Key Research and Development Program of China under Grant 2018YFC0831002, in part by the Key Research and Development Program of Shandong Province under Grant 2018GGX101045, and in part by the Natural Science Foundation of Shandong Province under Grant ZR2020MF132. *(Corresponding author: Zhe Li.)*

Sheng Fang, Kaiyu Li, and Zhe Li are with the College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266100, China (e-mail: fangs99@126.com; likyoo@sdust.edu.cn; lizhe@sdust.edu.cn).

Jinyuan Shao is with the Key Laboratory of Urban Environment and Health, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 364021, China, and also with the School of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: jyshao@iue.ac.cn).

Color versions of one or more figures in this letter are available at https://doi.org/10.1109/LGRS.2021.3056416.

Digital Object Identifier 10.1109/LGRS.2021.3056416

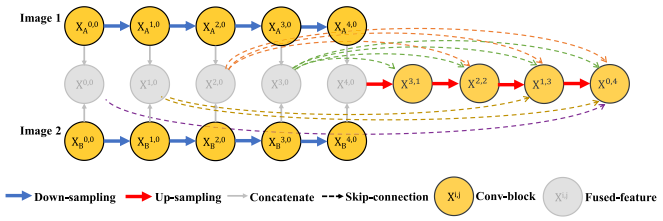

Fig. 2. Schematic of information transfer between the encoder and a sub-decoder of SNUNet-CD.

connected siamese network for change detection. Through dense skip connections between encoder and decoder, and between decoder and decoder, it can maintain high-resolution, fine-grained representations. Since our final backbone network structure is a combination of siamese network and NestedUNet, we name it SNUNet-CD.

The backbone of SNUNet-CD has multiple groups of outputs at different levels. To implement the natural aggregation of final low-level and high-level features, and suppress semantic gaps of deep supervision, we introduce Channel Attention Module (CAM) [17] in deep supervision and propose the Ensemble Channel Attention Module (ECAM). Experimental results prove that ECAM can aggregate and refine features of multiple semantic levels and get the better results.

The contributions of this letter are mainly as follows:

1) We propose a densely connected SNUNet-CD based on NestedUNet for RS change detection. SNUNet-CD alleviates the loss of localization information in deep layers of neural network.
2) The ECAM is proposed to aggregate and refine features of multiple semantic levels, which suppress semantic gaps and localization error in some extent.
3) Through a series of experimental comparisons, our method is superior to other state-of-the-art (SOTA) methods in F1-Score and computational complexity. Our source code is released at https://github.com/likyoo/Siam-NestedUNet.

This letter is organized as follows. Section II describes the change detection method proposed in this letter. Section III contains a series of quantitative comparisons and analyses through experiments. Finally, the conclusion of this letter is drawn in Section IV.

II. Methodology

A. Network Architecture

SNUNet-CD is a standard encoder–decoder architecture and uses the siamese network as encoder. Bi-temporal images are input into two branches of siamese network respectively, and parameters are shared between the two branches. In this way, the same convolution filters are used to extract the features of two images, and the same position in the feature map is activated. Due to the siamese network separately extracts bi-temporal image features, the concatenation is used to fuse the features between two siamese branches to ensure the integrity of the information.

In order to maintain high-resolution features and fine-grained localization information, we use the dense skip connection mechanism between the encoder and decoder, as shown in Fig. 2. The features of two branches are fused during "Image 1" and "Image 2" are down-sampling, and the fused high-resolution, fine-grained features are successively transmitted to the decoder through skip connections, compensating for the loss of position information in the deep layers of the decoder. For example, after the first down-sampling, the outputs of $\boldsymbol{X}_A^{1,0}$ and $\boldsymbol{X}_B^{1,0}$ are obtained, and the fused feature $\boldsymbol{X}^{1,0}$ is obtained by concatenating them. When dense skip connections, first, the fused feature $\boldsymbol{X}^{1,0}$ is concatenated with the $2\times$ up-sampled output of $\boldsymbol{X}^{2,2}$ and input into the $\boldsymbol{X}^{1,3}$ convolution block; second, since the $2\times$ up-sampled output of $\boldsymbol{X}^{1,3}$ is twice the size of $\boldsymbol{X}^{1,0}$, an up-sampling operation is needed for $\boldsymbol{X}^{1,0}$. Similarly, in order to achieve dense connections for the entire backbone, up-sampling is needed for every node in the encoder except the original size node.

Therefore, the complete flowchart of SNUNet-CD is shown in Fig. 3. "Image 1" and "Image 2" are input into the siamese encoder network, and for the output of each node after down-sampling, there will be a sub-decoder to restore it to the original size. The fine localization features of the encoder are transmitted to four sub-decoders through skip connections, which means the shallow localization information is directly applied to the deep layers, so that the fine-grained information can be maintained. Furthermore, in order to further utilize the fine features, the output of the nodes in shallower sub-decoders will be connected to the nodes of deeper sub-decoders of the same size.

In Fig. 3(a) each node $\boldsymbol{X}^{i,j}$ denotes a convolution block, as shown in Fig. 3(c). The convolution block is designed as a residual unit structure [18]. But in particular, the shortcut connection is after the first convolution layer to maintain the unity of all convolution blocks.

Let $\boldsymbol{x}^{i,j}$ denotes the output of node $\boldsymbol{X}^{i,j}$, $\boldsymbol{x}^{i,j}$ is formulated as follows:

$$\boldsymbol{x}^{i,j} = \begin{cases} \mathcal{P}(\mathcal{H}(\boldsymbol{x}^{i-1,j})) & j = 0 \\ \mathcal{H}\left(\left[\boldsymbol{x}_A^{i,0}, \boldsymbol{x}_B^{i,0}, \ \mathcal{U}(\boldsymbol{x}^{i+1,j-1})\right]\right) & j = 1 \\ \mathcal{H}\left(\left[\boldsymbol{x}_A^{i,0}, \boldsymbol{x}_B^{i,0}, [\boldsymbol{x}^{i,k}]_{k=1}^{j-1}, \ \mathcal{U}\left(\boldsymbol{x}^{i+1,j-1}\right)\right]\right) & j > 1 \end{cases} \quad (1)$$

The function $\mathcal{H}(\cdot)$ denotes the operation of the convolution block. The function $\mathcal{P}(\cdot)$ denotes a 2×2 max pooling operation for down-sampling. The function $\mathcal{U}(\cdot)$ denotes up-sampling using transpose convolution. [] denotes the concatenation on the channel dimension and aims to fuse features. When $j = 0$, the encoder down-samples and extracts features; when $j > 0$, the dense skip connection mechanism starts to work, and the fine-grained features in the encoder are successively transmitted to the deep decoder.

In detail, the number of channels of the feature map in SNUNet-CD gradually increases with the deepening of the encoder, and gradually decreases with the deepening of the decoder. Fig. 4 shows the change in the number of channels.

B. ECAM

The backbone of SNUNet-CD finally has four outputs of the same size as the original images, the outputs $\{\boldsymbol{x}^{0,j}, j \in \{1, 2, 3, 4\}\}$ of nodes $\{\boldsymbol{X}^{0,j}, j \in \{1, 2, 3, 4\}\}$. Although these

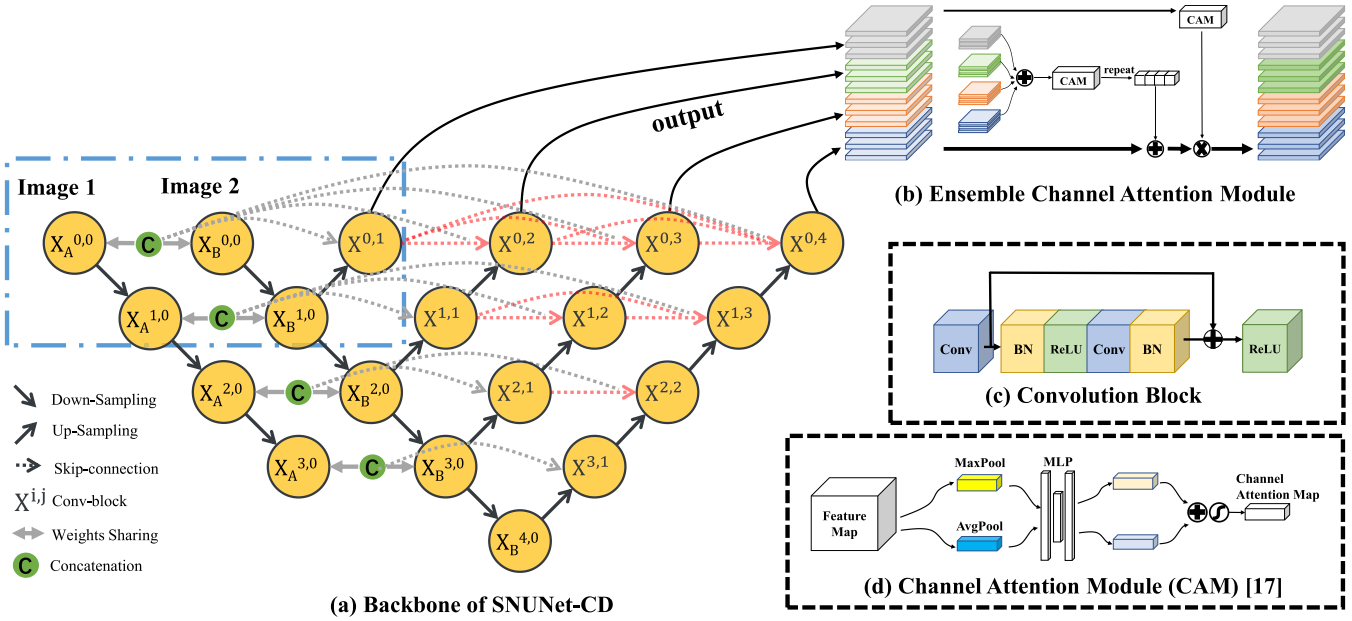

Fig. 3. Illustration of the proposed architecture. (a) is the backbone of SNUNet-CD, downward arrows and upward arrows indicate down-sampling and up-sampling respectively, the dotted arrows indicate skip connections for transmitting fine-grained features (gray indicates connections between encoder and sub-decoders, and red indicates connections between sub-decoders and sub-decoders); the node $X^{i,j}$ indicates a convolution block, the detailed structure is shown in (c). (b) is our ECAM. (d) is a regular CAM from CBAM [17].

four groups of feature maps are in the same size, they have different semantic levels and spatial position representations. Specifically, the outputs of shallow sub-decoder have finer-grained features and more precise localization, and the outputs of the deep sub-decoder have more coarse-grained features and richer semantics. Intuitively, when fusing these features, an automatic channel selection strategy is needed to suppress semantic gaps and localization differences.

Therefore, as shown in Fig. 3(b), the ECAM is designed to automatically select and focus on more effective information between different groups. Structurally, ECAM is a natural expansion of CAM [17] in deep supervision and ensemble learning. First, four groups of output in the backbone are summed, and a CAM is followed to extract the intra-group relation. Synchronously, four groups of output are concatenated, and another CAM is followed to extract the inter-group relation. To be specific, ECAM can be formulated as follows:

$$\boldsymbol{CAM}(F) = \sigma\left(\boldsymbol{MLP}(\boldsymbol{AvgPool}(\boldsymbol{F})) + \boldsymbol{MLP}(\boldsymbol{MaxPool}(\boldsymbol{F}))\right) \tag{2}$$

$$\boldsymbol{M}_{intra} = \boldsymbol{CAM}(\boldsymbol{x}^{0,1} + \boldsymbol{x}^{0,2} + \boldsymbol{x}^{0,3} + \boldsymbol{x}^{0,4}) \tag{3}$$

$$\boldsymbol{F}_{ensemble} = [\boldsymbol{x}^{0,1}, \boldsymbol{x}^{0,2}, \boldsymbol{x}^{0,3}, \boldsymbol{x}^{0,4}] \tag{4}$$

$$\boldsymbol{M}_{inter} = \boldsymbol{CAM}(\boldsymbol{F}_{ensemble}) \tag{5}$$

$$\begin{aligned}&\boldsymbol{ECAM}(\boldsymbol{F}_{ensemble})\\ &\quad = (\boldsymbol{F}_{ensemble} + \boldsymbol{repeat}_{(4)}(\boldsymbol{M}_{intra})) \otimes \boldsymbol{M}_{inter}\end{aligned} \tag{6}$$

where σ denotes the sigmoid function, MLP denotes a multilayer perception, AvgPool and MaxPool denote average pooling and max pooling operations respectively. [] denotes the concatenation of groups of feature maps, the function $\boldsymbol{repeat}_{(n)}(\cdot)$ denotes the operation of repeating the attention map $\boldsymbol{n}$ times and concatenating in the channel dimension, $\otimes$ denotes element-wise product. Finally, a 1×1 convolution is

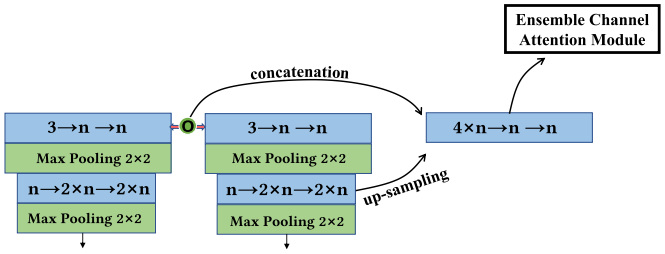

Fig. 4. Changes of the number of channels in SNUNet-CD. This part is indicated in Fig. 3(a) by a blue dotted frame. "3" indicates 3 channels RGB image, "n" indicates the initial number of channels of feature map.

followed to get the change map $\hat{\boldsymbol{Y}}$:

$$\hat{\boldsymbol{Y}} = \hbar(\boldsymbol{ECAM}(\boldsymbol{F}_{ensemble})) \tag{7}$$

where $\hbar(\cdot)$ denotes a 1×1 convolutional layer to generate a $2 \times H \times W$ change map $\hat{\boldsymbol{Y}}$ ("2" indicates change and no change).

C. Details of Loss Function

In the field of change detection, the number of unchanged pixels is often far more than the number of changed pixels. To weaken the impact of sample imbalance, we use a hybrid loss function (the combination of weighted cross-entropy loss and dice loss), which is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{wce}} + \mathcal{L}_{\text{dice}}. \tag{8}$$

To describe the weighted cross-entropy loss, the change map $\hat{\boldsymbol{Y}}$ is regarded as a set of points, which can be represented as:

$$\hat{\boldsymbol{Y}} = \{\hat{\boldsymbol{y}}_{\boldsymbol{k}}, \boldsymbol{k} = 1, 2, \ldots, H \times W\} \tag{9}$$

where $\hat{\boldsymbol{y}}_{\boldsymbol{k}}$ denotes a point in $\hat{\boldsymbol{Y}}$, and it contains two values here. $\boldsymbol{H}$ and $\boldsymbol{W}$ denote the height and width of $\hat{\boldsymbol{Y}}$, which is the

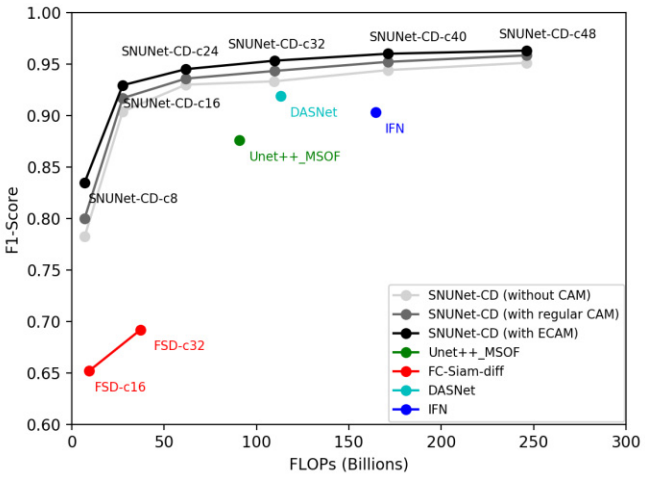

Fig. 5. FLOPs versus F1-Score. "-cn" means the initial number of channels is n.

same size as the original images. The weighted cross-entropy (WCE) loss can be formulated as:

$$\mathcal{L}_{\text{wce}} = \frac{1}{H \times W} \sum_{k=1}^{H \times W} \text{weight[class]} \cdot \left(\log \left(\frac{\exp(\hat{\boldsymbol{y}}[k][\text{class}])}{\sum_{l=0}^{1} \exp(\hat{\boldsymbol{y}}[k][l])} \right) \right) \quad (10)$$

where the value of "class" is 0 or 1, corresponding to the unchanged and changed pixels, respectively. At the same time, the change map $\hat{\boldsymbol{Y}}$ participates in calculating the dice loss after a Softmax layer:

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2 \cdot \boldsymbol{Y} \cdot \text{softmax}(\hat{\boldsymbol{Y}})}{\boldsymbol{Y} + \text{softmax}(\hat{\boldsymbol{Y}})} \quad (11)$$

where $\boldsymbol{Y}$ indicates the ground truth.

III. Experiments

A. Data Set and Evaluation Metrics

To evaluate our method, we design a series of comparative experiments on CDD [1] data set, one of the most common evaluation data sets in the field of change detection. CDD data set contains 11 pairs of multispectral (consist of R, G, B) images taken in different seasons obtained by Google Earth (Digital Globe), and the spatial resolution of these images is 3 cm/px to 100 cm/px. In [1], 16 000 pairs of images with a size of 256 × 256 pixels are generated from the original image through cropping and rotation operations. And 10 000 image pairs for the training set, 3000 for validation and testing sets respectively. For quantitative metrics evaluation, three indicators are used: Precision, Recall, and F1-Score.

B. Implementation Details

We implement SNUNet-CD using the Pytorch framework. In the training process, the batch size is set to 16, and Adam is applied as an optimizer. Learning rate is set to 1 e–3 and decays by 0.5 every 8 epochs. The weights of each convolutional layer are initialized by the KaiMing normalization. We conduct experiments on a single NVIDIA Tesla v100 and train for 100 epochs to make the model converge.

TABLE I
Performance Comparison on CDD Data Set

Method / Channel	Params(M)	Precision	Recall	F1
FC-EF [5] / 16 *	1.35	0.609	0.583	0.592
FC-Siam-conc [5]/16 *	1.54	0.709	0.603	0.637
FC-Siam-diff [5] / 16 *	1.35	0.762	0.573	0.652
FC-Siam-diff [5] / 32 *	5.39	0.783	0.626	0.692
Unet++_MSOF[12]/32	11.00	0.895	0.871	0.876
IFN [9] / -	35.72	0.950	0.861	0.903
DASNet [11] / -	16.25	0.914	0.925	0.919
SNUNet-CD (our) / 8	**0.75**	**0.893**	**0.784**	**0.834**
SNUNet-CD (our) / 16	**3.01**	**0.943**	**0.916**	**0.929**
SNUNet-CD (our) / 24	**6.77**	**0.951**	**0.938**	**0.944**
SNUNet-CD (our) / 32	**12.03**	**0.956**	**0.949**	**0.953**
SNUNet-CD (our) / 40	**18.80**	**0.961**	**0.959**	**0.960**
SNUNet-CD (our) / 48	**27.06**	**0.963**	**0.962**	**0.962**

The symbol "*" means our re-implemented results.

C. Comparison and Analysis

Several representative deep learning-based methods in the field of change detection are selected to compare. FC-EF, FC-Siam-conc, and FC-Siam-diff [5] are baseline models for change detection, they are simple combinations of UNet and siamese network. Unet++_MSOF [12] inputs concatenated images into UNet++ backbone and uses the multiple side-outputs fusion for deep supervision. IFN [9] uses pretrained vgg16 as encoder and proposes a multiscale deep supervision strategy. DASNet [11] introduces a dual attention mechanism [19] to the contrastive method in change detection.

In our comparative experiments, we consider the tradeoff between accuracy evaluation and the floating point of operations (FLOPs). And in SNUNet-CD, the network width (the number of channels of the feature map, indicated by "n" in Fig. 4) is regarded as an adjustable hyper-parameter to balance accuracy and FLOPs. As we all know, the wider network requires more calculations, and captures more fine-grained features [20].

The quantitative results listed in Table I suggest that our SNUNet-CD can get excellent performance in change detection of complex scenarios. On CDD data set, the 16-channel SNUNet-CD can achieve better performance than other SOTA change detection methods, and it only has 3.01 M parameters and only 27.56 G FLOP. In fact, the dense connections between the siamese encoders and the decoders are the main factors that bring improvement, as demonstrated by the results of SNUNet-CD (without CAM) in Fig. 5. The transmission of fine-grained localization information improves the edge determination of changed objects and the detection of small targets, as shown in Fig. 6. On the other hand, different from Unet++_MSOF, the application of the siamese network makes SNUNet-CD avoid the entanglement and noncorrespondence of spatial features in bi-temporal images.

As shown in Fig. 5, the effect of ECAM cannot be ignored. Although the dense connection mechanism alleviates the loss of localization information in the deep layers, there is still the semantic gap when using deep supervision. ECAM can refine

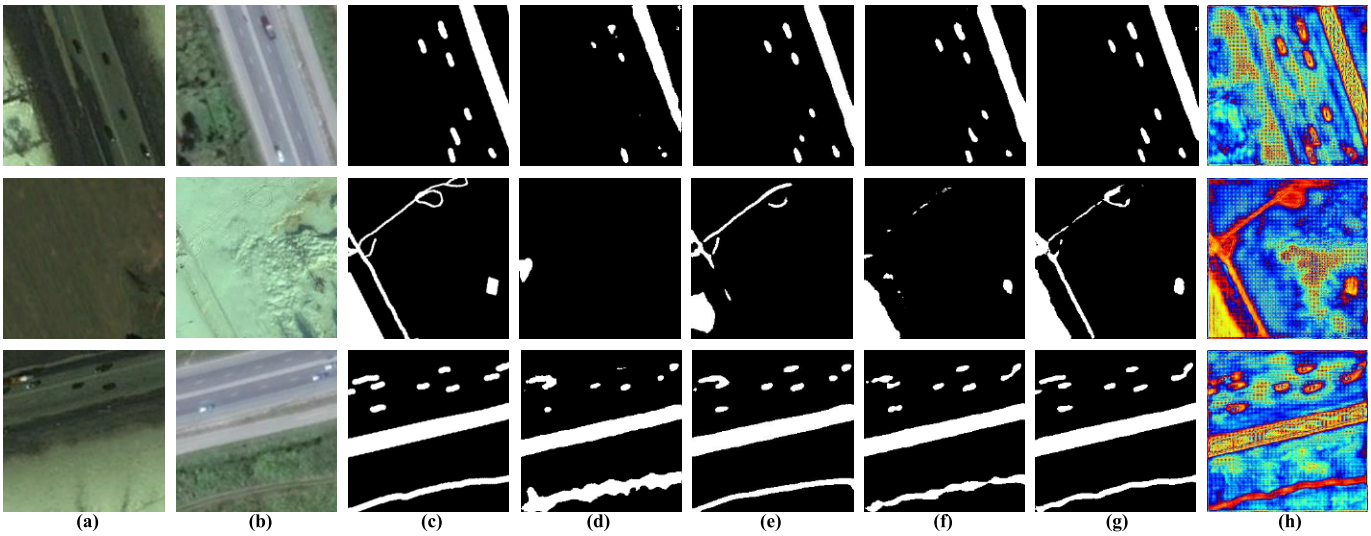

Fig. 6. Visualization results on CDD data set: (a) and (b) are original bi-temporal images; (c) is the ground truth; (d) is the result of FC-Siam-diff; (e) is the result of IFN; (f) is the result of DASNet; (g) is our result of SNUNet-CD with 32 channels; (h) is the heatmap after ECAM in SNUNet-CD.

intra-group and inter-group relations from features of different granularities and semantic levels, alleviating the semantic gap to some extent. As shown in Fig. 6(h), we use the heatmap to visualize feature maps after ECAM. The heatmap is obtained from the variance of all feature maps. Through ECAM, the changed area gets higher energy, and the energy at the edge of the target is stronger. In other words, the edge of the changed target is strengthened and localized more accurately, which improves the detection performance.

IV. Conclusion

In this letter, we propose a densely connected siamese network for change detection of very high resolution (VHR) images, namely SNUNet-CD. Through dense skip connections between encoder and decoder, and between decoder and decoder, SNUNet-CD can maintain high-resolution, fine-grained representations, and alleviate the uncertainty of the pixels at the edge of the changed target and the determination miss of small targets. Structurally, SNUNet-CD can be regarded as the combination of siamese network, NestedUNet, and ECAM. Compared to other researches on change detection, SNUNet-CD can detect more detailed changes and better balance the accuracy evaluation and FLOPs. In the future, we will focus on the extraction and utilization of spectral information in change detection and transfer our methods to the field of hyperspectral RS images.

References

[1] M. A. Lebedev, Y. V. Vizilter, O. V. Vygolov, V. A. Knyaz, and A. Y. Rubis, "Change detection in remote sensing images using conditional adversarial networks," *ISPRS Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. XLII-2, pp. 565–571, May 2018.

[2] W. A. Malila, "Change vector analysis: An approach for detecting forest changes with Landsat," in *Proc. LARS Symposia*, 1980, p. 385.

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

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Cham, Switzerland: Springer, 2015, pp. 234–241.

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

[6] F. Rahman *et al.*, "Siamese network with multi-level features for patch-based change detection in satellite imagery," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Nov. 2018, pp. 958–962.

[7] 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.

[8] E. Guo *et al.*, "Learning to measure change: Fully convolutional siamese metric networks for scene change detection," 2018, *arXiv:1810.09111*. [Online]. Available: http://arxiv.org/abs/1810.09111

[9] C. Zhang *et al.*, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 183–200, Aug. 2020.

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

[11] J. Chen *et al.*, "DASNet: Dual attentive fully convolutional siamese networks for change detection of high resolution satellite images," 2020, *arXiv:2003.03608*. [Online]. Available: http://arxiv.org/abs/2003.03608

[12] D. Peng, Y. Zhang, and H. Guan, "End-to-end change detection for high resolution satellite images using improved UNet++," *Remote Sens.*, vol. 11, no. 11, p. 1382, 2019.

[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[14] Z. Zhou *et al.*, "UNet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 3–11.

[15] Z. Zhou *et al.*, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Dec. 2019.

[16] G. Huang, M. M. R. Siddiquee, M. Tajbakhsh, and J. Liang, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.

[17] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2016, pp. 630–645.

[19] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[20] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016, *arXiv:1605.07146*. [Online]. Available: http://arxiv.org/abs/1605.07146