Intelligent-BCD: A Novel Knowledge-Transfer Building Change Detection Framework for High-Resolution Remote Sensing Imagery

Haiming Zhang, Guorui Ma, and Yongxian Zhang

***Abstract*—In the study of human social development and ecological environment monitoring, building change detection (BCD) is essential. Rapid and accurate BCD in complicated scenes and multiview high-resolution (HR) remote sensing images have garnered a lot of attention with the continual increase of image resolution and diverse satellite imaging modes. However, conventional BCD methods almost exclusively focus on the performance under specific, standard datasets and do not consider the robustness under off-nadir imaging and complex scene conditions. To address these issues, we propose an adaptative knowledge-driven multilevel attention BCD framework, called intelligent-BCD, in which we innovatively deploy four attention mechanisms. The compatibility design of the channel and spatial attention mechanism enables the model to fully mine deep features and meet BCD tasks in multiple imaging modes and complex scenes. Meanwhile, we propose a weight pool concept, a knowledge accumulation way of modeling attention adaptive. The model parameters obtained from the training of different data are stored to form the weight pool, and the weight for transfer application is reused in the new BCD task. Furthermore, we propose a new loss function, the dynamic domain loss function, that successfully addresses the problem of sample imbalance while motivating the model to pay greater attention to challenging sample learning. The benchmark experimental results for the four datasets, namely LEVIR-CD, LEVIR-CD+, WHU Building Dataset, and S2Looking, show that intelligent-BCD is superior to the competing methods in quantitative and qualitative evaluation. The transfer experiment in MtS-WH dataset demonstrates that intelligent-BCD has an excellent generalization and robustness.**

***Index Terms*—Attention, building change detection (BCD), knowledge-driven, remote sensing, transfer-learning.**

I. Introduction

CHANGE detection (CD), one of the basic technologies in the field of remote sensing, analyzes changes in the location and status of objects in the area using remote sensing images recorded in the same region at various periods [1]. Buildings are the most common man-made structures, and they are an important part of land use, urban landscape, regional environments, etc. Building CD (BCD) is a popular study issue in many CDs because timely and reliable information about building changes is critical in catastrophe assessment, urban expansion, and environmental change studies [2].

Sensor and optoelectronic technology advancements have significantly aided remote sensing research advancement. Remote sensing image resolution has improved from medium- and low- to high- or even very high resolution, and satellite sensors have entered the multiview and multimodal imaging era. As a result, the CD will be able to use a greater quantity and variety of remote sensing data, which will open up new possibilities. For example, it allows CD to detect changes in instances (e.g., buildings) on the micro-scale of the Earth's surface. Some early researchers made BCD attempts using medium- and low-resolution imagery [3]. However, the image resolution hampered their efforts, resulting in only coarse BCD outputs. Pixel-based and object-based are two types of classical approaches, such as image difference [4], image ratio [5], change vector analysis [6], regression analysis [7], slow feature analysis [8], multivariate alteration detection [9], etc., have performed different CD tasks based on medium- and low-resolution imagery, but BCD works are rare. Many old methods are becoming challenging to employ as high-resolution (HR) remote sensing photos become the primary data source for BCD because many approaches produce salt and pepper noise or struggle to deal with the challenges provided by complex spectral and textural properties [10].

Using various convolutional neural networks (CNNs) for CD has become the mainstream of study with the introduction of deep learning [11] technology. Utilizing CNN's hierarchical feature representation, we can extract high-level features from HR images and use them in downstream processes to enable end-to-end product production using the data-driven paradigm [10], [12]. Hence, this meets most jobs based on remote sensing images, such as object detection [13], scene classification [14], semantic segmentation [15], CD [12], etc. Many CNNs-based models have been designed and employed in different CD tasks, such as FC-EF-Res [16], ChangeNet [17], ReCNN [18], Siam-CRNN [19], SEW-Net [20], SCDNET [21], etc. Some of these models have complex structures, while others include novel mechanisms, all with the goal of achieving faster convergence and greater accuracy. Furthermore, attention mechanisms [10], transformer mechanisms [22], generative-adversarial mechanisms [23], and others are widely used to accurately find instances of small objects (e.g., buildings) in HR remote sensing

Manuscript received May 18, 2022; revised June 8, 2022; accepted June 15, 2022. Date of publication June 17, 2022; date of current version June 30, 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1004603. *(Corresponding author: Guorui Ma.)*

The authors are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: zhanghaiming@whu.edu.cn; mgr@whu.edu.cn; zhyx009@whu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3184298

images with rich, detailed information. These mechanisms play a significant role in the BCD task and achieve good effectiveness. We discovered, however, that the studies described above are conducted under particular, standard data settings. That is, the imaging must be done in on-nadir mode, and the image must have undergone stringent correction and registration, with the additional constraint that the scene in the image is not too intricate. While images in the actual world are diverse and complicated, using only one or two standard datasets to demonstrate the method's feasibility is plainly not rigorous.

High-level products that have undergone thorough preprocessing and are imaged at near-nadir viewing angles are widely used images in public databases. Furthermore, because the up-front error accumulation of the entire pipeline is modest, poor illumination or seasonal fluctuations in the image provide a suitable precondition for the model. Remarkably, no generalized method can conduct a BCD job with the same precision on other datasets, and transferring a method developed for one dataset to other datasets for application is difficult. As a result, it's critical to provide a compatible BCD framework that can extract image feature information effectively, is suitable for multi-view imaging images, and is transferable.

In this article, we propose a novel BCD framework called intelligent-BCD. This framework is an adaptive, knowledge-driven BCD approach based on a multilevel attention mechanism. The contributions of this article are as follows:

1) Intelligent-BCD is compatible with four attention strategies and is a pioneering attempt at an attention mechanism. It can perform effective BCD work in both on-nadir and off-nadir remote sensing images and can effectively extract key feature information in complex images.
2) We propose a new concept of weight pool, which is unprecedented in CD research. The weight parameters obtained by training in different datasets are continuously accumulated and stored as knowledge, and the method is perfectly transferred through adaptive weight search in a new dataset.
3) We propose a new loss function called dynamic domain loss function (DDLF). This function is very useful for unbalanced training samples and difficult sample feature mining. Combined with the channel and spatial attention synergy mechanism we designed in the framework, the perfect expression of spatial image information and better retention of feature information is achieved.
4) We conducted experiments on four publicly available datasets and achieved satisfactory results. To verify the transferability of the framework, we independently annotated a reference change dataset based on MtS-WH dataset and published it. The result is available at https://github.com/Haiming-Z/MtS-WH-reference-map.

II. Materials

A. Benchmark Datasets

To validate the proposed method completely, we used four types of datasets in the benchmark experiment (see Fig. 1), i.e., the LEVIR-CD [24], LEVIR-CD+, WHU Building dataset [25], and S2Looking [26].

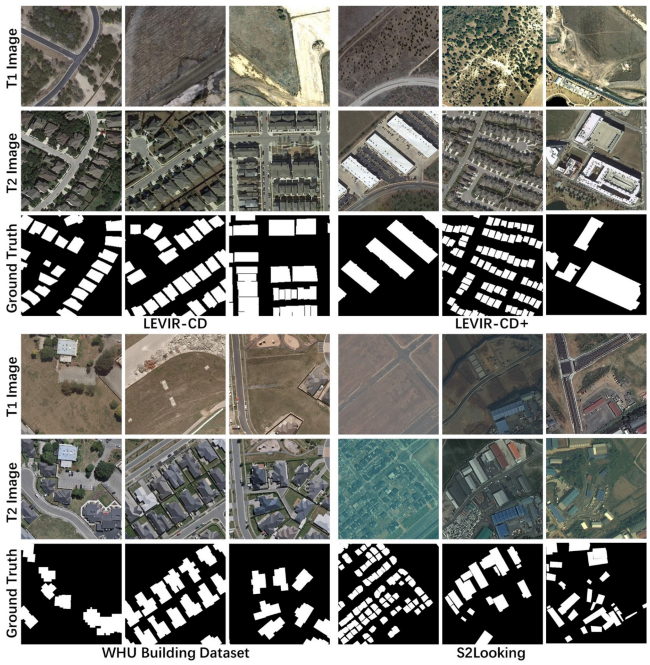

Fig. 1. Some examples of the benchmark datasets.

LEVIR-CD is composed of Google Earth images with a resolution of 0.5 m, including a total of 637 image patch pairs with a size of 1024 × 1024. These bi-temporal remote sensing images with a time span of 5–14 years were obtained from 20 different regions that sit in several cities in Texas, USA. The image contains large amounts of various building instances, 31333 in total. And LEVIR-CD+ is based on LEVIR-CD. Its image's acquisition time varies from 2002 to 2020. It has a total of 985 images, including about 80000 building instances.

The WHU Building Dataset contains two aerial images with 0.075 m spatial resolution, covering Christchurch, New Zealand, acquired in 2012 (with 12796 buildings) and 2016 (with 16077 buildings), respectively. There was a 6.3-magnitude earthquake in February 2011, which destroyed a lot of buildings. The size of the image is 10065 × 11645, and there is a corresponding reference map. In addition, the image coverage is 20.5 square kilometers, which is valuable for validating the robustness of the model to handle large-scale continuous imagery.

S2Looking is a special kind of dataset. It contains large-scale side-looking satellite images captured at various off-nadir angles, obtained with the Rolling Imaging Mode of Optical Satellite. The dataset contains 5000 bitemporal image pairs of rural areas and over 65920 annotated instances of changes around the world. The resolution of the image is 0.5–0.8 m, and the size is 1024 × 1024. The illumination variation in the image is greater, and the scene is more complex. This is more challenging for the model's performance, applicability, and stability.

B. Transfer Application Dataset

If a proven method can still achieve good results when using a completely new type of data, then it is certain that this method has practical value. After validating our approach with

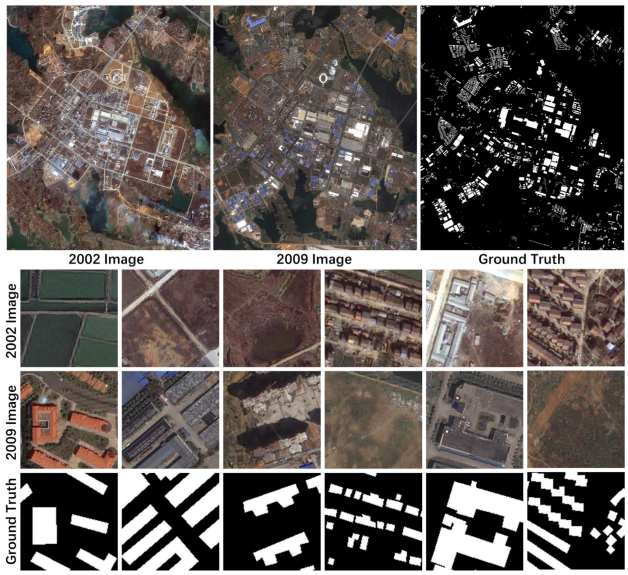

Fig. 2. MtS-WH dataset and some scenes examples.

Benchmark datasets, we apply our approach to a new dataset that we annotated based on the MtS-WH dataset (see Fig. 2).

The MtS-WH dataset consists of two large-scale HR remote sensing images with a size of 7200 × 6000 obtained by the IKONOS sensor. The images were acquired in February 2002 and June 2009, respectively, with a resolution of 1 m. The coverage of the image is Hanyang District, Wuhan City, which is developing rapidly in China. The image includes a variety of changing scenes, including changes in buildings in rural areas, cities, industrial areas, and residential areas. Moreover, there is a certain off-nadir angle in the image. We annotated these types of scene changes and published our results.

III. Methodology

A. Intelligent-BCD: The Adaptative Knowledge-Driven Multilevel Attention BCD Framework

The schematic diagram of the overall architecture of our proposed Intelligent-BCD is shown in Fig 3. The four attention mechanisms in intelligent-BCD are responsible for different tasks. Channel attention, spatial attention, and weight pool together to form a new multilevel attention network called MLA-Net. It fully and efficiently models the image features at the channel and spatial level and adapts to different data based on the weight pool. The DDLF is responsible for guiding MLA-Net to focus on the information mining of difficult samples to obtain the optimal training effect.

B. Basic Architecture of MLA-Net

BCD work favors more sensitive methods to the spectral, textural, and structural features of a specific man-made object. Therefore, it is necessary to use new mechanisms and characteristics to analyze, simulate, and construct models for brain-like perception and cognition, both at the macroscopic and microscopic levels.

We believe that the neural network should not only be the weight adjustment in the current sense, but more importantly, its structure should have variability, plasticity, learning, and dynamics, which are currently ignored. Based on this insight, we designed MLA-Net (see Fig. 4). MLA-Net is a lightweight network with a simple structure and few parameters. Because the multiplicative impact during gradient back-propagation may create unstable weight updates as the network deepens, resulting in inferior performance on CD tasks [27]. The backbone of MLA-Net is linear, with traditional encoding and decoding branches, and both input and output are distinct. The transmission of mistakes can be considerably reduced with this end-to-end strategy. We employed dilated convolution in both the encoding and decoding branches, as well as batch normalization and ReLU operations, to minimize the loss of internal data structure and hierarchical spatial information. The bitemporal image pairs are concatenated into the network, and the output terminal generates a binary map.

We used skip connections in the network, and the difference is that we embed the spatial attention mechanism (SAM) in it. Before concatenation operations, all SAMs are run, which can effectively combine low- and high-level features, facilitate cross-layer contextual information transmission, get more compact global information, and direct the network to focus on task-relevant BCD areas. The convolution operation processes the image layer by layer and transmits information, uses the information collected from the receptive field to form a high-dimensional output, and then obtains a feature map through a nonlinear activation function. If the feature map of the lth layer in the encoding path is $\boldsymbol{x^l}$. SAM will use a gating vector g to guide the model to focus on the target region of $\boldsymbol{x^l}$. And weaken the feature response of the background region. This process generates an attention coefficient α, which is multiplied by the corresponding element of $\boldsymbol{x^l}$. To obtain a spatial attention output that suppresses the response of the background region and strengthens the response of the target region, SAM is formulated as follows:

$$q_{att}^{l} = \psi^{T} \left(\sigma_1 \left(W_x^T x_i^l + W_g^T + b_g\right)\right) + b_\psi \tag{1}$$

$$\alpha_i^l = \sigma_2 \left(q_{att}^l \left(x_i^l,\ g_i;\ \Theta_{att}\right)\right) \tag{2}$$

where σ_1 is the ReLU activation function, $\sigma_2 = \frac{1}{1+e^{-x_i}}$ correspond to the sigmoid activation function, Θ_{att} represents a set of parameters of SAM, namely linear transformations W_x, W_g, ψ, and bias terms b_ψ, b_g.

SAM better expresses spatial details and aggregates multilevel contextual information. However, the parsing of changing scenes in images also requires the network to model the interdependencies between channels adequately. It needs the network to be more directional, capable of recalibrating features, emphasizing important features, and suppressing secondary features. Therefore, we used the channel attention mechanism (CAM) in MLA-Net and combined it organically with SAM. CAM is a global information modeling approach for fusing information in the local receptive field that overcomes the limitations of convolution operations [28]. We used CAM in both the encoding and decoding branches, making it work after each convolution

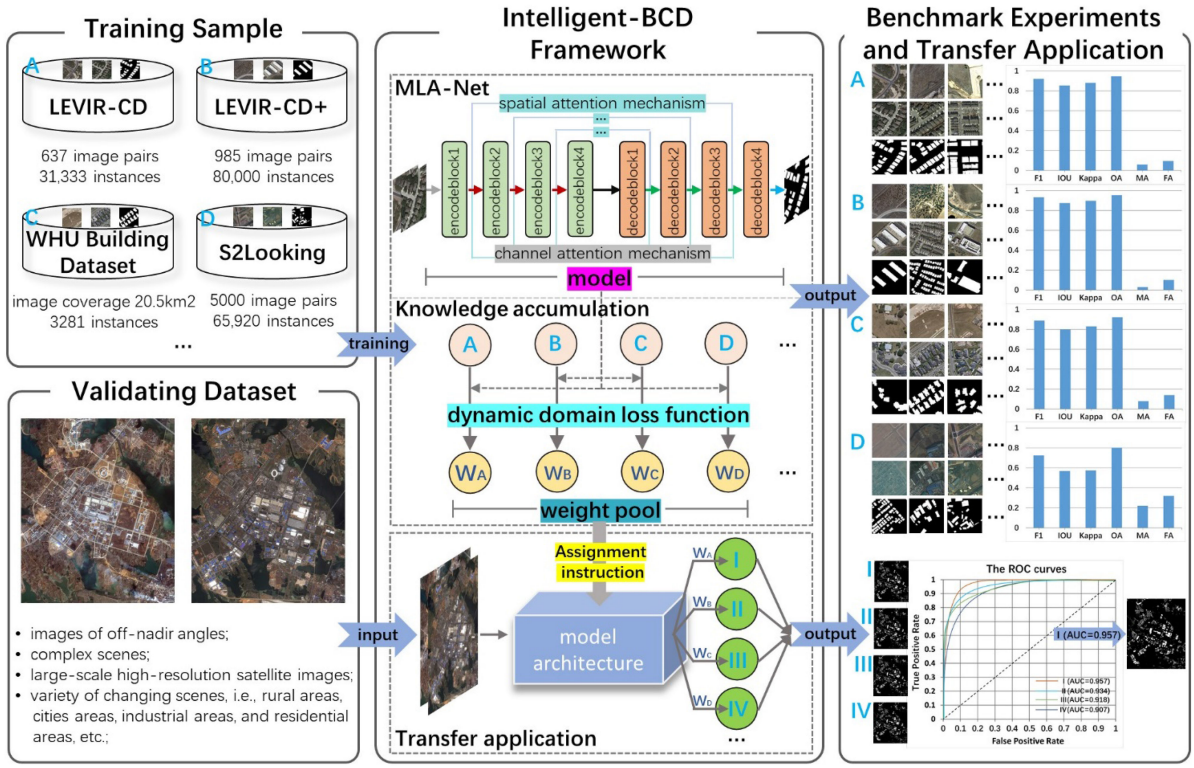

Fig. 3. Overview of intelligent-BCD framework.

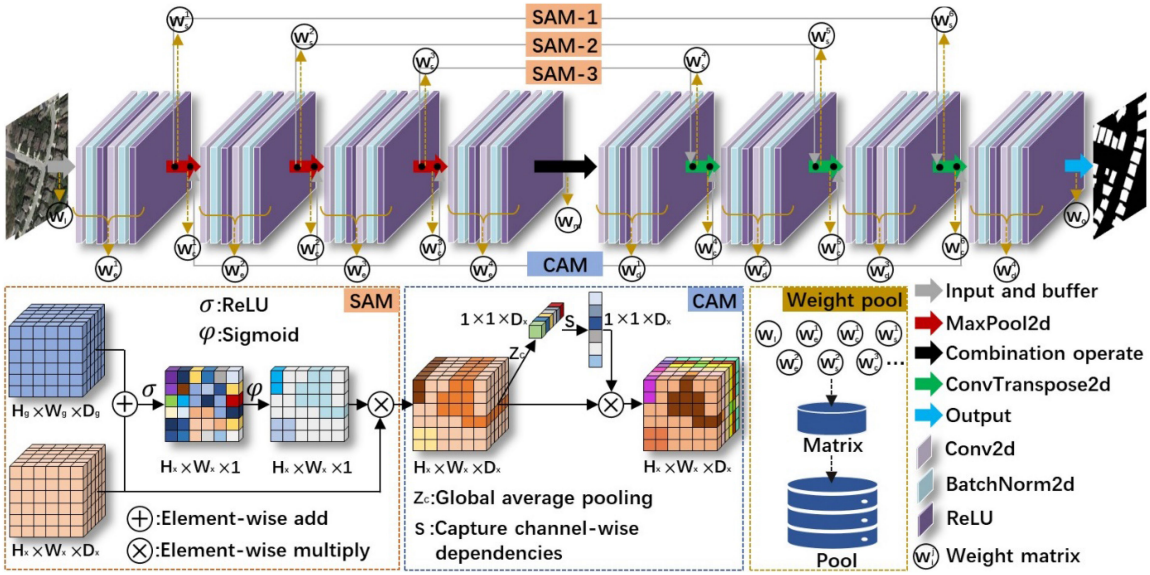

Fig. 4. Architecture of the MLA-Net.

operation, explicitly modeling the dynamic nonlinear relationship between channels, redistributing weights to each channel, enhancing important information, and suppressing secondary information. The output of the SAM will be used as the input of the CAM into the downstream information representation of the spatial relationship of the pixels. The compatible use of CAM and SAM simulates the cognitive laws of the human brain and can fully understand the image. If the output of a convolution is $X \in \mathbb{R}^{H\times W\times C}$, CAM can generate a feature descriptor s of size $1\times1\times C$, and X is multiplied by s to get the channel-weighted output. CAM is formulated as follows:

$$z_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} x_c(i,j) \tag{3}$$

$$s = \sigma\left(g(z,W)\right) = \sigma\left(W_2\delta\left(W_1 z\right)\right) \tag{4}$$

$$\tilde{X} = s_c \times x_c \tag{5}$$

where x_c is the cth feature map, $(i,\ j)$ is the pixel position in the feature map, z is the global description, δ is the ReLU activation function, W_1, and W_2 are the weight matrices of the two fully connected layers, σ is the Sigmoid function, and s_c is the cth global descriptor.

According to our statistics, there are at least dozens of neural network models in BCD research. It is worth mentioning that the solution to the problem is by no means a simple adjustment process of parameters. Simply seeking the optimal global solution under specific data cannot solve various practical problems, nor is it the original intention of artificial intelligence. The continuous accumulation and iteration of models is a process of constant evolution, and it is also a process of constantly enriching ways to solve problems. Based on this insight, we proposed a new concept of weight pool, which aims to endow the network model with the ability to accumulate knowledge and achieve adaptive applications in different data. If MLA-Net completes model training on A, B, C, and D data, and the obtained weight matrix corresponding to the entire network is W_A, W_B, W_C, and W_D, then the weight pool $P = \{W_A, W_B, W_C, W_D\}$. When a new BCD task needs to be performed, the model does not need to be trained redundantly, and the "knowledge" accumulated in *P* can be used for transfer application, and the BCD result can be obtained quickly. It's worth mentioning that MLA-Net has a straightforward and easy-to-maintain structure. It can be trained with more data, and the weights can be continuously accumulated to obtain a more diverse knowledge base that can saves a lot of time training the model and reuses valuable training resources. We will also establish an open community where everyone can upload their basic network structure and training results or use others' networks to train under new data and share new training results. In this way, we want to continuously accumulate process assets to achieve wider application and information sharing of the network model. This is our community website.[1]

C. New Dynamic Domain Loss Function

BCD is a binary classification task. Its goal is to distinguish between buildings (positive samples) and nonbuilding (negative samples). This is reflected in the ability of the model to learn positive and negative samples during the training process. Because the true percentage of positive and negative samples is not balanced, the learning difficulty likewise varies. Some works [29], [30] have studied the case where there are more negative samples than positive samples. And the latter, in this case, is more challenging to learn. These works have achieved satisfactory results. However, we believe that the proportion of the two samples changes during the training process, and the model should be dynamically guided to learn the critical information of positive and negative samples instead of always ignoring negative samples and strengthening positive samples. There are more positive samples than negative samples, especially when the model is fed data in fixed-size batches. Therefore, we proposed a DDLF to adjust the distribution of positive and negative sample loss values to properly train the model in complicated scenarios and obtain more robust detection capabilities. It's formulated as follows:

$$L_{DD} = \begin{cases} -\mu_\alpha (1-p_t)^\theta \log p_t, \alpha < \beta \ or \ \frac{\beta}{\alpha+\beta} \geq 0.25 \\ -\mu_\beta (1-p_t)^\theta \log p_t, \alpha > \beta \ and \ \frac{\beta}{\alpha+\beta} < 0.25 \end{cases} \tag{6}$$

where p is the probability predicted by the model ($p_t = p$ or $p_t = 1-p$), μ is the class weight, θ is the sample adjustment factor (default = 2), α and β, respectively, represent the number of positive and negative samples in each batch. $\mu_\alpha = 0.25$ when $\alpha < \beta$ or $\frac{\beta}{\alpha+\beta} \geq 0.25$, $\mu_\beta = \frac{\beta}{\alpha+\beta}$ when $\alpha > \beta$ and $\frac{\beta}{\alpha+\beta} < 0.25$.

D. Comparative Approaches and Evaluation Metrics

We use six traditional semantic segmentation networks as comparative approaches to assess intelligent-BCD's performance, i.e., FCN, UNet, SegNet, UNet++, Attention U-Net, and DeepLabv3+.We use six evaluation metrics to evaluate our method quantitatively namely, F1-Score (F_1), Intersection-over-Union (IoU), Kappa coefficient (Kappa), overall accuracy (OA), missing alarm (MA), and false alarm (FA). Their calculation formulas are as follows:

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{7}$$

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{8}$$

$$M = \frac{(FN + TP)(FP + TP) + (FN + FT)(TN + FP)}{FP + TP + TN + FN} \tag{9}$$

$$\text{Kappa} = \frac{TN + TP - M}{TN + FP + TP + FN - M} \tag{10}$$

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

$$\text{MA} = \frac{FN}{TP + FN} \tag{12}$$

$$\text{FA} = \frac{FP}{TP + FP} \tag{13}$$

where TP denotes the number of correctly identified positive samples, FP denotes the number of wrongly classified negative samples, FN denotes the number of incorrectly classed positive samples, and TN denotes the number of correctly classified negative samples. M is the conversion factor.

IV. Experimental Results and Discussion

A. Experimental Settings

The intelligent-BCD is powered by a workstation with an Intel(R) Xeon(R) W-2245 CPU (3.90GHz, 64 GB RAM) and a single NVIDIA GeForce RTX 3090, which runs PyTorch with python3.7 as the backend. Each dataset consists of three subdatasets: training, validation, and test, where the patch size is 256×256 pixels. Table I provides more information. The Adam optimizer was chosen as the optimizer, with a learning rate of 0.001. All of the models were trained for 50 epochs, and the batch sizes for training and validation are 32 and 16, respectively.

[1][Online]. Available: https://github.com/Haiming-Z/Weight-sharing-community-.

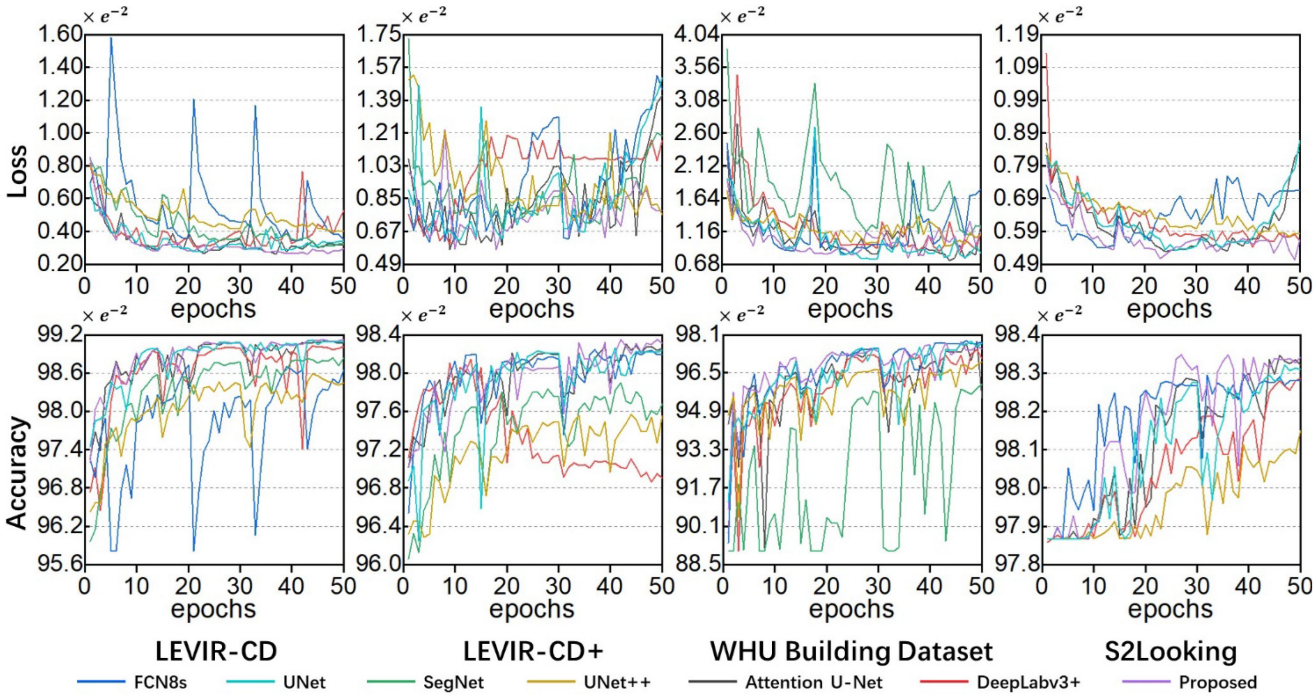

Fig. 5. Curves of various models on four datasets with iterations.

TABLE I
STATISTICS OF THE DATASETS SPLIT

Split	LEVIR-CD	LEVIR-CD+	WHU Building Dataset	S2Looking
Training	7120	10,192	6500	3500
Validation	1024	3168	500	500
Test	2048	2400	434	1000

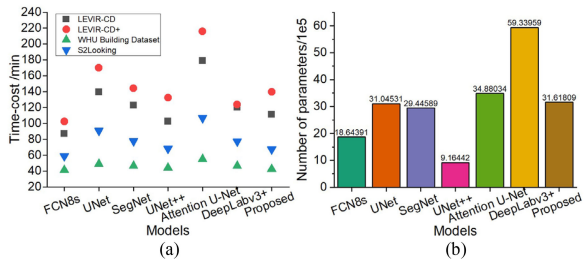

Fig. 6. Comparisons of different models in terms of (a) time-cost and (b) model size.

B. BCD Results and Performance Analysis

We trained each of these seven models using the benchmark datasets and recorded the accuracy and loss values when the models were trained (see Fig. 5). Even more to the point, the DDLF is involved in the training of all models.

The results show that the MLA-Net converges faster and stabilizes almost within 30 epochs. That may be attributed to the network's lightweight structure and efficient attention mechanism. The convergence process of other models takes a long time and is accompanied by large fluctuations. However, all models basically achieved convergence within 50 epochs. A plausible explanation for this phenomenon seems to point entirely to the existence of a DDLF. During training, loss and accuracy values are the most intuitive and essential reflections of model performance. Therefore, it can be preliminarily concluded that the internal design principle of MLA-Net is reasonable and robust. The DDLF is critical to the model's training process.

The time required for each model to complete 50 iterations over four datasets and the number of model parameters are also of interest. Because if a disaster occurs in an area, fast BCD is the guarantee for follow-up work. Meanwhile, the model's usability improves as the amount of model parameters decreases. As shown in Fig. 6, we visualized the model's training time and the number of parameters.

The time consumption of MLA-Net is tolerable, as shown in Fig. 6(a). Its performance in time consumption is almost only inferior to FCN8s and DeepLabv3+, which is undoubtedly related to the use of pre-trained models in both. The pre-trained model is used as the front-end network, which dramatically reduces the time cost of parameter adjustment. The embedding of SAM and CAM in MLA-Net does not impose much computational burden on the model. In contrast, the time consumption of Attention U-Net is almost always the highest, probably because it only adds an attention mechanism to UNet without considering the complexity of the overall network. Furthermore, the amount of parameters of MLA-Net is moderate. UNet++ has the least number of parameters due to the pruning strategy used. FCN8s also has a smaller number of parameters because it does not use the fully connected layer of vgg16 although it uses vgg16 as the front-end network. The number of parameters of UNet, SegNet, Attention-U-Net, and MLA-Net are not very different and belong to the same order of magnitude. It can be assumed that MLA-Ne does not increase its number of parameters substantially while obtaining better performance. This also indicates that it achieves a good value balance. It is worth mentioning

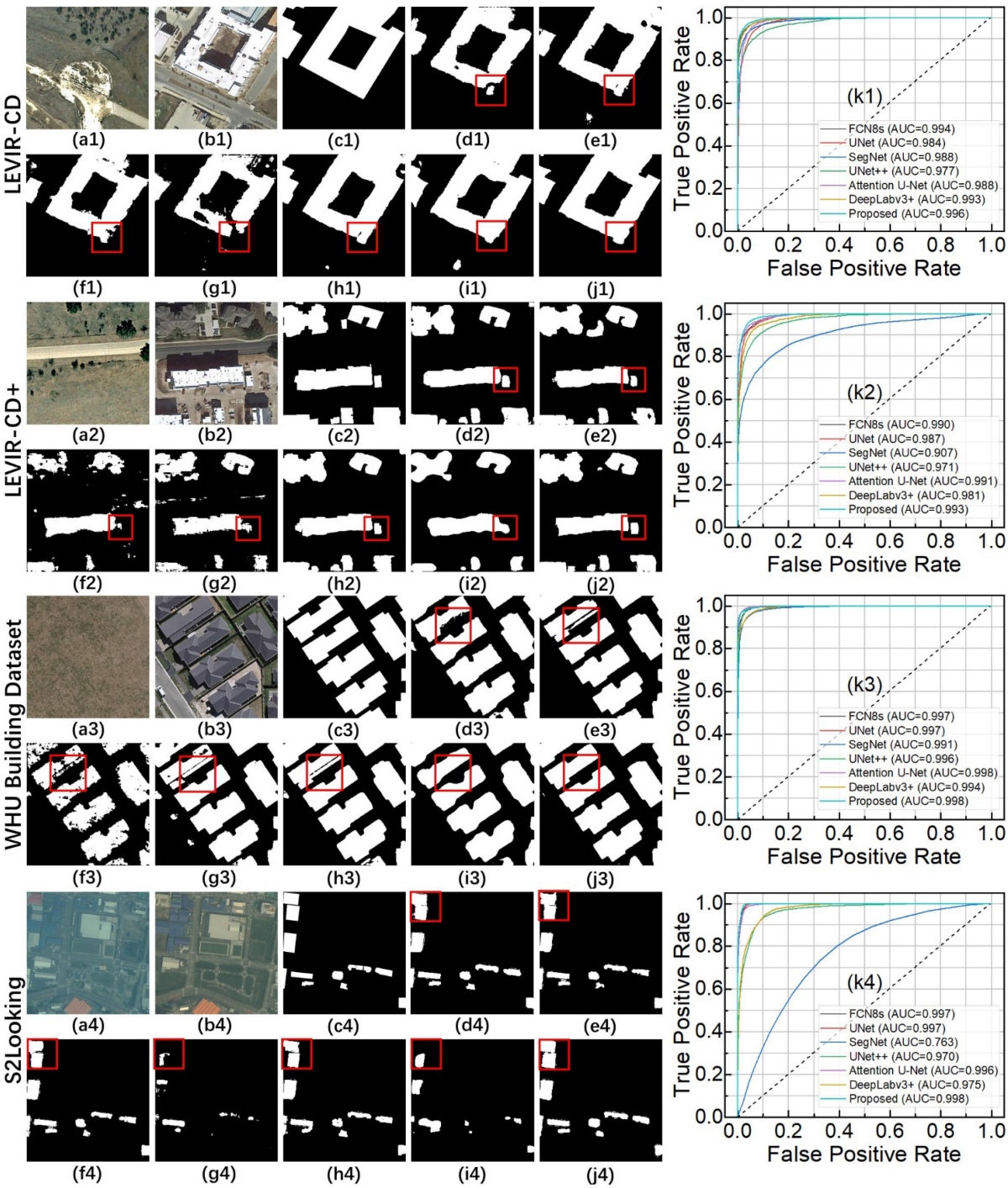

Fig. 7. Visual performance of different models on the benchmark test datasets. (a) T1 image. (b) T2 image. (c) Ground truth. (d) FCN8s. (e) UNet. (f) SegNet. (g) UNet++. (h) Attention U-Net. (i) DeepLabv3+. (j) Proposed. (k) ROC curves.

that the pre-trained model in DeepLabv3+ greatly increases the number of parameters, which undoubtedly reduces the usability of the model.

We chose several exemplary examples for visualization and showed the receiver operating characteristic (ROC) curves to evaluate the model performance intuitively, as shown in Fig. 7.

We calculated quantitative evaluation metrics for all model prediction results on the test datasets, and the results are reported in Tables II and III. From the statistics, it is not difficult to find that MLA-Net leads almost all evaluation metrics. On the four datasets, the mean values of F_1, IoU, Kappa, and OA reached 0.818, 0.726, 0.784, and 0.953, respectively. We stopped the model's training at only 50 epochs, and the number of patches in the test dataset is plentiful. In this case, the four metrics above fully illustrate the excellent performance of MLA-Net. Furthermore, the F_1 value performs the best, which shows that MLA-Net considers both the detection accuracy and the completeness of target extraction. We note that the MA and

TABLE II
QUANTITATIVE RESULTS OF SEVERAL APPROACHES ON THE LEVIR-CD AND LEVIR-CD+ DATASETS

Method	LEVIR-CD						LEVIR-CD+				
	F_1	IoU	Kappa	OA	MA	FA	F_1	IoU	Kappa	OA	MA
FCN8s	0.752	0.642	0.740	0.964	0.305	0.075	0.735	0.621	0.709	0.952	0.323
UNet	0.836	0.750	0.824	0.975	0.204	0.068	0.753	0.646	0.729	0.957	0.295
SegNet	0.800	0.695	0.780	0.966	0.265	0.070	0.691	0.572	0.660	0.942	0.361
UNet++	0.732	0.620	0.712	0.963	0.329	0.141	0.603	0.483	0.574	0.941	0.455
Attention U-Net	0.840	0.755	0.827	0.976	0.204	0.065	0.763	0.657	0.739	0.957	0.283
DeepLabv3+	0.822	0.727	0.807	0.972	0.228	0.070	0.747	0.633	0.721	0.953	0.301
Proposed	0.846	0.761	0.832	0.976	0.201	0.062	0.775	0.671	0.751	0.957	0.255

TABLE III
QUANTITATIVE RESULTS OF SEVERAL APPROACHES ON THE S2 LOOKING AND WHU BUILDING DATASETS

Method	WHU Building Dataset						S2Looking				
	F_1	IoU	Kappa	OA	MA	FA	F_1	IoU	Kappa	OA	MA
FCN8s	0.858	0.794	0.801	0.930	0.174	0.053	0.769	0.674	0.730	0.940	0.259
UNet	0.849	0.782	0.789	0.934	0.196	0.036	0.716	0.598	0.671	0.923	0.374
SegNet	0.749	0.645	0.669	0.894	0.295	0.110	0.710	0.585	0.653	0.904	0.389
UNet++	0.783	0.695	0.719	0.911	0.239	0.143	0.748	0.648	0.713	0.938	0.293
Attention U-Net	0.846	0.778	0.786	0.931	0.200	0.029	0.774	0.681	0.744	0.942	0.300
DeepLabv3+	0.791	0.701	0.726	0.908	0.276	0.047	0.761	0.651	0.723	0.944	0.316
Proposed	0.868	0.799	0.811	0.939	0.178	0.031	0.781	0.673	0.743	0.940	0.311

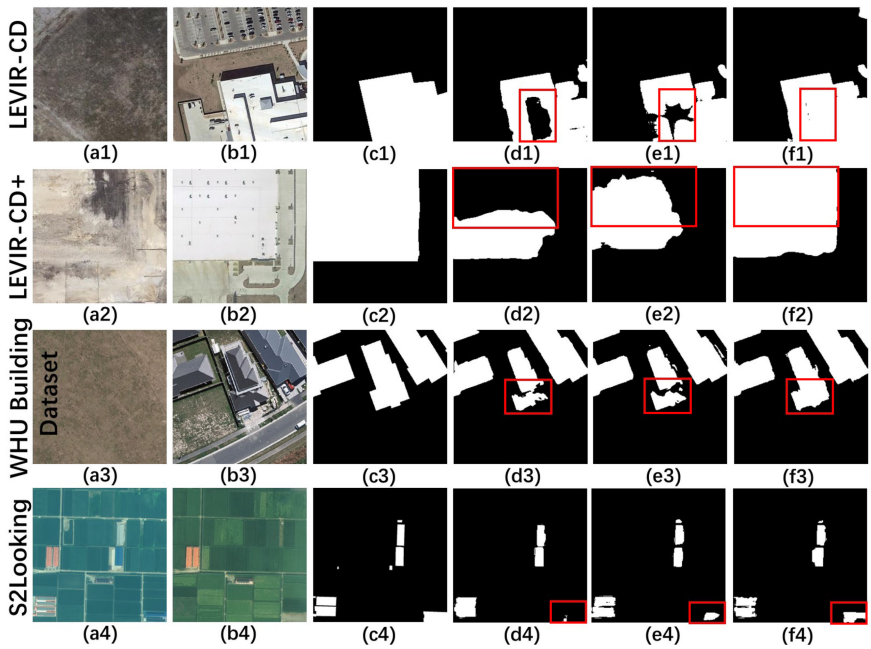

Fig. 8. Visual validation results of attentional mechanisms and DDLF. (a) T1 image. (b) T2 image. (c) Ground truth. (d) Baseline method. (e) Baseline with attention mechanism (Baseline-AM). (f) Baseline with attention mechanism and DDLF (Baseline-AM&DDLF).

FA of MLA-Ne are low, which indicates that it can precisely localize the changed regions with the help of the joint attention mechanism. And some models without attention mechanisms perform poorly on MA and FA metrics. The visual results show that MLA-Net has the best performance and outperforms the other models substantially. Its detection results have clear and accurate edges and high building integrity (marked by the red box in Fig. 7.). It is worth mentioning that MLA-Net is very robust to the interference factors in the image. That is, there are few false detection spots in the detection results. The ROC curve reflects the performance of the model. MLA-Net has the best operational characteristics and the most significant area under the curve (AUC) value. That shows that MLA-Net has the best overall performance for the prediction probability of each pixel in the image and the most robust generalization ability.

C. Ablation Study

To further evaluate the efficiency of the attention mechanism and DDLF, we conducted ablation experiments on four datasets. In this session, MLA-Net without the attention mechanism and the DDLF (CrossEntropyLoss instead) is used as the baseline method. In Fig. 8, some samples are visualized. Fig. 9. compares the evaluation metrics of different approaches.

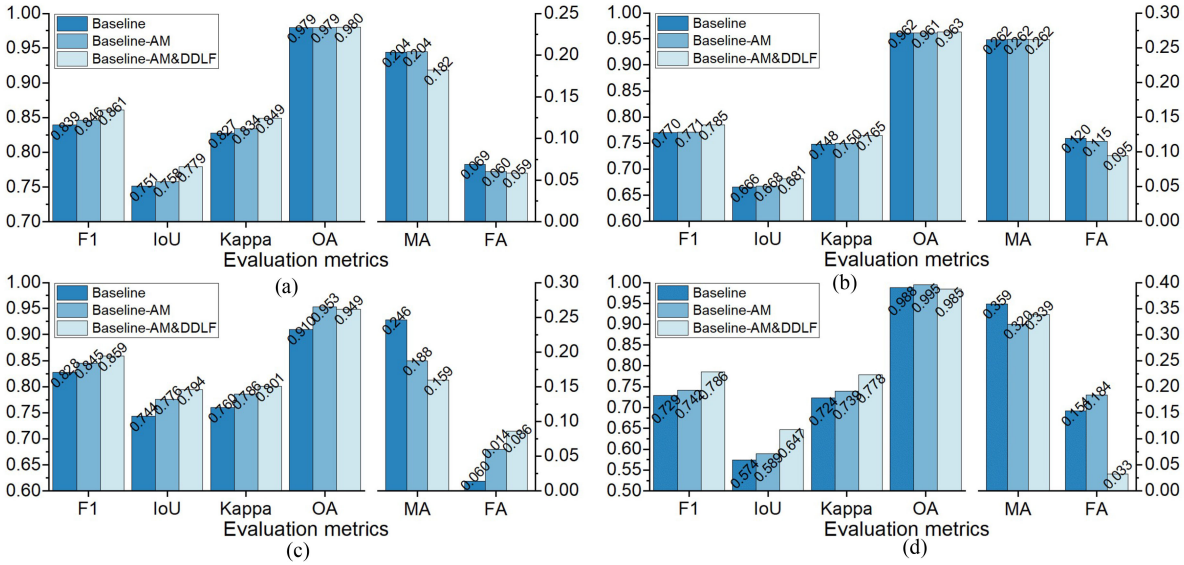

Fig. 9. Evaluation metrics value of attention mechanism and DDLF. (a) LEVIR-CD. (b) LEVIR-CD+. (c) WHU Building Dataset. (d) S2Looking.

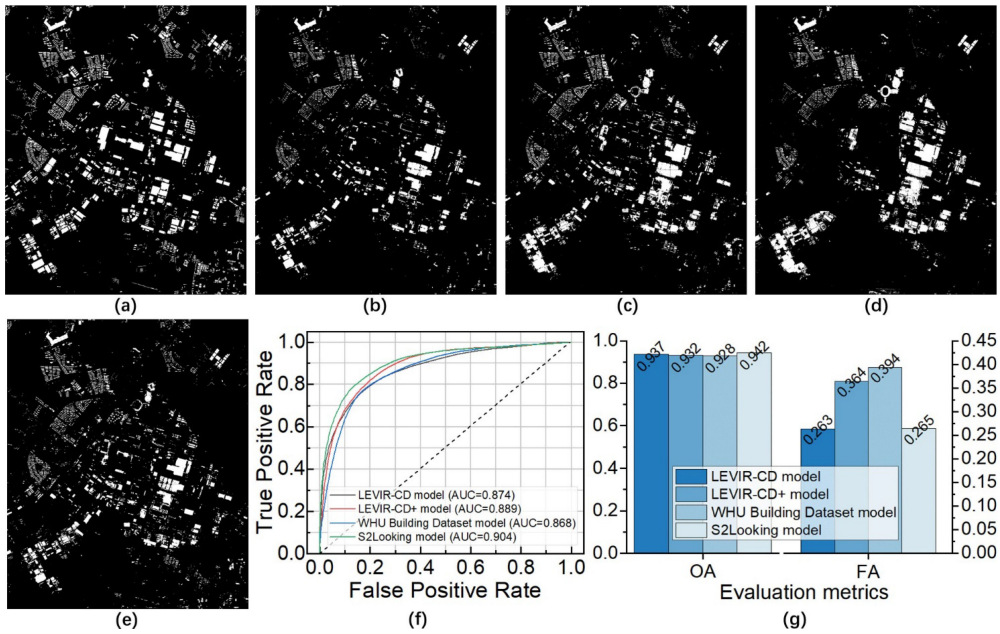

Fig. 10. Results of transfer application. (a) Ground truth. (b) LEVIR-CD model. (c) LEVIR-CD+ model. (d) WHU building dataset model. (e) S2Looking model. (f) ROC curves. (g) Evaluation histogram.

The use of the attention mechanism and DDLF bring a significant performance boost to the model. The detection of buildings by the baseline model is always incomplete, and even some changed buildings are not detected (see Fig. 8 red box). That demonstrates how the attention mechanism directs the model's attention to learning the changing regions of the building and increases the model's sensitivity to them. The DDLF guides the model to learn positive and negative samples discriminately. Combining the two makes the model's detection of buildings entirely and accurately. It also shows that the compatibility between the two is better. The maximum increase in $\boldsymbol{F}_1$, IoU, Kappa, and OA are 5.7, 7.35, 5.4, and 4.3%, respectively. This quantitative evaluation result shows that using the attention mechanism and DDLF improves model performance considerably.

D. Analysis of Transfer Application

To verify the usability of our proposed weight pool concept, We transfer the MLA-Net model trained on the benchmark dataset to the MtS-WH dataset. Notably, we train four models within 50 epochs using only 100 patches. According to our

observation, the training time of each model is no more than 8 min, and the validation accuracy is higher than 85%, which is outstanding in the fast BCD task. All models predict the entire image, and the prediction results, ROC curves (for 100 patches), and evaluation histogram (metrics of OA and FA) are shown in Fig. 10.

After training with a modest number of training samples, the models trained on diverse datasets may accomplish perfect transfer applications, which satisfies the work requirements of efficient BCD. We counted the OA and FA values of all models. The data reflects that all models have high detection accuracy on new data (OA values are all higher than 92%), and the lowest false detection is 26.3%. We noticed that the S2Looking model was more adaptable to unfamiliar images (AUC value of 0.904). This may be because the imagery in this dataset was taken under an off-nadir imaging angle. The uniqueness of the data places more demands on the model, making the trained model more resilient, as well as its capacity to deal with complex scenes and multiview images. Qualitatively and quantitatively, our proposed concept of weight pool is feasible and can be used even in broader BCD tasks. The results of the experiments demonstrate that intelligent-BCD will have a distinct application value in BCD tasks.

V. Conclusion and Future Work

In this article, we propose a novel BCD framework named intelligent-BCD. This framework is a comprehensive BCD pipeline that includes a new attention module, a new concept of weight pool, and a new loss function. We design an attention mechanism that combines spatial and channel mechanisms to efficiently utilize the effective features and strengthen the model's perception ability. Models trained on different datasets have different data characteristics. Our proposed concept of weight pool is an interesting attempt at multitask application of different models. The experimental results prove the correctness of our attempt. Considering the imbalance of training samples and the difficulty of mining stubborn sample information, we designed a new loss function named DDLF. The introduction of this loss function motivates the model to focus on the learning of crucial samples. And to a certain extent, the generalization ability and the convergence speed of the model are improved. The efficacy and robustness of MLA-Net are confirmed by experimental findings on benchmark datasets, which highlight the importance of the attention mechanism and DDLF. The quantitative and qualitative evaluation results on the transfer dataset show that the innovative idea of weight pool is feasible and has substantial promotion value. The BCD tasks of the whole process comprehensively reflect that intelligent-BCD is robust, viable, and efficient for BCD under complex and diverse data.

In the future, we will use different datasets for more extensive model training to obtain more diverse benchmark models. At the same time, optimize the model structure and try to apply intelligent-BCD to other tasks.

References

[1] X. Huang, Y. Cao, and J. Li, "An automatic change detection method for monitoring newly constructed building areas using time-series multi-view high-resolution optical satellite images," *Remote Sens. Environ.*, vol. 244, 2020, Art. no. 111802.

[2] A. Asokan and J. Anitha, "Change detection techniques for remote sensing applications: A survey," *Earth Sci. Inform.*, vol. 12, pp. 143–160, 2019.

[3] F. Yamazaki and M. Matsuoka, "Remote sensing technologies in post-disaster damage assessment," *J. Earthq. Tsunami*, vol. 1, pp. 193–210, 2007.

[4] N. A. Quarmby and J. L. Cushnie, "Monitoring urban land cover changes at the urban fringe from SPOT HRV imagery in South-East England," *Int. J. Remote Sens.*, vol. 10, pp. 953–963, 2010.

[5] P. J. Howarth and G. M. Wickware, "Procedures for change detection using Landsat digital data," *Int. J. Remote Sens.*, vol. 2, pp. 277–291, 2007.

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

[7] A. K. Ludeke, R. C. Maggio, and L. M. Reid, "An analysis of anthropogenic deforestation using logistic regression and GIS," *J. Environ. Manage.*, vol. 31, no. 3, pp. 247–259, 1990.

[8] C. Wu, B. Du, and L. Zhang, "Slow feature analysis for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2858–2874, May 2014.

[9] B. Du, Y. Wang, C. Wu, and L. Zhang, "Unsupervised scene change detection via latent dirichlet allocation and multivariate alteration detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 12, pp. 4676–4689, Dec. 2018.

[10] Q. Ding, Z. Shao, X. Huang, and O. Altan, "DSA-Net: A novel deeply supervised attention-guided network for building change detection in high-resolution remote sensing images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 105, 2021, Art. no. 102591.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[12] M. Wang, H. Zhang, W. Sun, S. Li, F. Wang, and G. Yang, "A coarse-to-fine deep learning based land use change detection method for high-resolution remote sensing images," *Remote Sens.*, vol. 12, 2020, Art. no. 1933.

[13] E. Liu, Y. Zheng, B. Pan, X. Xu, and Z. Shi, "DCL-Net: Augmenting the capability of classification and localization for remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7933–7944, Jan. 2021.

[14] Q. Zhu, Y. Zhong, L. Zhang, and D. Li, "Scene classification based on the fully sparse semantic topic model," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5525–5538, Oct. 2017.

[15] L. Huan, X. Zheng, S. Tang, and J. Gong, "Learning deep cross-scale feature propagation for indoor semantic segmentation," *ISPRS J. Photogram. Remote Sens.*, vol. 176, pp. 42–53, 2021.

[16] R. Caye Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Multitask learning for large-scale semantic change detection," *Comput. Vis. Image Understanding*, vol. 187, 2019, Art. no. 102783.

[17] A. Varghese, J. Gubbi, A. Ramaswamy, and P. Balamuralidhar, "ChangeNet: A deep learning architecture for visual change detection," in *Proc. Eur. Conf. Comput. Vis.*, 2019, vol. 11130, pp. 129–145.

[18] L. Mou, L. Bruzzone, and X. X. Zhu, "Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 924–935, Nov. 2018.

[19] H. Chen, C. Wu, B. Du, L. Zhang, and L. Wang, "Change detection in multisource VHR images via deep siamese convolutional multiple-layers recurrent neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2848–2864, Dec. 2020.

[20] H. Zhang *et al.*, "A novel squeeze-and-excitation W-Net for 2D and 3D building change detection with multi-source and multi-feature remote sensing data," *Remote Sens.*, vol. 13, 2021, Art. no. 440.

[21] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, and P. He, "SCDNET: A novel convolutional network for semantic change detection in high resolution optical remote sensing imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 103, 2021, Art. no. 102465.

[22] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 5607514, doi: 10.1109/TGRS.2021.3095166.

[23] W. Zhao, X. Chen, X. Ge, and J. Chen, "Using adversarial network for multiple change detection in bitemporal remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Nov. 2022, Art. no. 8003605, doi: 10.1109/LGRS.2020.3035780.

[24] 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, 2020, Art. no. 1662.

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

[26] L. Shen *et al.*, "S2Looking: A satellite side-looking dataset for building change detection," *Remote Sens.*, vol. 13, 2021, Art. no. 24.
[27] J. Liu, M. Gong, A. K. Qin, and K. C. Tan, "Bipartite differential neural network for unsupervised image change detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 3, pp. 876–890, Mar. 2020.
[28] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
[29] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2999–3007.
[30] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 811–815, May 2021.

Haiming Zhang received the B.S. degree in surveying and mapping engineering from the Jilin Jianzhu University, Changchun, China, in 2018, the M.E. degree in surveying and mapping engineering from Jilin University, Changchun, China, in 2021. He is currently working toward the Ph.D. degree in photogrammetry and remote sensing with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China.

His research interests include high-resolution remote sensing image change detection, land cover and land use change analysis, building damage detection and assessment, and disaster analysis.

Guorui Ma received the B.S. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2002, and the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2007.

He is a Research Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. He has long been engaged in basic research and technology development of photogrammetry and remote sensing applications in the fields of target assurance, public security counter-terrorism, and disaster emergency response. He has authored or coauthored more than 30 academic papers in remote sensing image target detection and recognition, semantic segmentation and extraction, disaster/damage assessment, big data prediction and early warning, participated/edited 5 monographs, authorized 18 national invention patents, 4 software copyrights

He is the recipient of one first prize of scientific and technological progress of the Ministry of Education and one first prize of scientific and technological progress of the army.

Yongxian Zhang received the B.S. degree in geographic information system from the Xinyang Normal University, Xinyang, China, in 2016, and the M.S. degree in photogrammetry and remote sensing from the Information Engineering University, Zhengzhou, China, in 2020. He is working toward the Ph.D. degree in photogrammetry and remote sensing with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, China.

His research interests include multi-modal remote sensing image matching & registration, and disaster damage assessment.