



Article

A Cross-Domain Change Detection Network Based on Instance Normalization

Yabin Song ¹, Jun Xiang ^{2,3,4,5} , Jiawei Jiang ^{3,4,6} , Enping Yan ^{3,4}, Wei Wei ^{2,5,*} and Dengkui Mo ^{3,4}

- ¹ Central South Academy of Inventory and Planning of NFGA, Changsha 410019, China; 20221100023@csuft.edu.cn
- ² Forestry Research Institute of Guangxi Zhuang Autonomous Region, Nanning 530002, China; xiangjun@csuft.edu.cn
- ³ Key Laboratory of State Forestry and Grassland Administration on Forest Resources Management and Monitoring in Southern Area, Changsha 410004, China; jiangjw26@mail2.sysu.edu.cn (J.J.); enpingyan@csuft.edu.cn (E.Y.); dengkuimo@csuft.edu.cn (D.M.)
- ⁴ College of Forestry, Central South University of Forestry & Technology, Changsha 410004, China
- ⁵ Guangxi Key Laboratory of Superior Timber Trees Resource Cultivation, Nanning 530002, China
- ⁶ School of Civil Engineering, Sun Yat-Sen University, Zhuhai 519082, China
- * Correspondence: 20180203@csuft.edu.cn

Abstract: Change detection is a crucial task in remote sensing that finds broad application in land resource planning, forest resource monitoring, natural disaster monitoring, and evaluation. In this paper, we propose a change detection model for cross-domain recognition, which we call CrossCDNet. Our model significantly improves the modeling ability of the change detection on one dataset and demonstrates good generalization on another dataset without any additional operations. To achieve this, we employ a Siamese neural network for change detection and design an IBNM (Instance Normalization and Batch Normalization Module) that utilizes instance normalization and batch normalization in order to serve as the encoder backbone in the Siamese neural network. The IBNM extracts feature maps for each layer, and the Siamese neural network fuses the feature maps of the two branches using a unique operation. Finally, a simple MLP decoder is used for end-to-end change detection. We train our model on the LEVIR-CD dataset and achieve competitive performance on the test set. In cross-domain dataset testing, CrossCDNet outperforms all the other compared models. Specifically, our model achieves an F1-score of 91.69% on the LEVIR-CD dataset and an F1-score of 77.09% on the WHU-CD dataset, where the training set was LEVIR-CD.

Keywords: change detection; deep learning; cross-domain; instance normalization



Citation: Song, Y.; Xiang, J.; Jiang, J.; Yan, E.; Wei, W.; Mo, D. A Cross-Domain Change Detection Network Based on Instance Normalization.

Remote Sens. **2023**, *15*, 5785.

<https://doi.org/10.3390/rs15245785>

Academic Editors: Peter M. Atkinson and Ce Zhang

Received: 7 November 2023

Revised: 11 December 2023

Accepted: 15 December 2023

Published: 18 December 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Convolutional neural networks (CNNs) based on deep learning have become a valuable tool in computer vision [1], enabling tasks such as image recognition [2], object detection [3–5], semantic segmentation [6–8], image super-resolution [9], and change detection [10,11]. Change detection is a crucial technique in remote sensing image analysis and is extensively used in land resource planning, forest resource monitoring [12,13], and natural disaster monitoring and assessment [14]. The goal of change detection is to identify and characterize surface changes from remote sensing data acquired at different times in a quantitative manner. This technique allows for semantic segmentation of the changed regions between the two different time periods, extending the scope of remote sensing analysis. Change detection involves a pixel-to-pixel visual interpretation task that enhances the quantitative analysis capability of the changed regions.

Deep learning models have been widely used in change detection tasks to extract semantic features at a deeper level. Some studies have utilized deep learning models that were originally designed for image segmentation tasks, such as constructing a Siamese encoder–decoder structure for change detection, which employs the U-Net++ [15] backbone

network. Furthermore, many researchers have aimed to improve the performance and efficiency in change detection by extracting deep semantic features, as shown in [16,17]. One noteworthy similarity among these change detection studies is their focus on exploring semantic features associated with deeper-level changes. The objective of semantic depth features is to juxtapose categories or characteristics of identical geographic features depicted in images captured across distinct time intervals. This analytical process seeks to identify alterations within these features, encompassing variations in building quantities, road expansion, or instances of deforestation. Expansive structures, notably buildings, frequently dominate substantial portions of the image. There is superior appropriateness regarding deep features for encapsulating semantic nuances within these extensive regions. Conversely, shallow features exhibit heightened effectiveness in addressing localized intricacies, comprising spatial elements like object edges and textures. These spatial features are employed to assess variations in the color, position, and shape of geographical elements across images acquired at different time points. Notable examples include structural translations or modifications of buildings and color shifts in forests resulting from seasonal variations.

Within the remote sensing community, significant distinctions in color domains among various remote sensing images arise due to seasonal and regional spatial variations. Consequently, it becomes imperative to comprehensively consider both semantic and spatial features to attain more precise and comprehensive results in change detection. This undertaking is of utmost importance in enhancing the cross-domain detection capabilities and the overall generalization of the change detection model. Figure 1 illustrates the significant differences in color between two change detection datasets. Although traditional change detection methods such as Image Difference, change vector analysis (CVA) [18], PCA, and K-means [19] can yield effective results in some simple scenarios, their performance significantly decreases when applied to other datasets for scene analysis. Most contemporary supervised remote sensing (RS) image change detection approaches are customized for equal-resolution bitemporal images. Conventional change detection models fail to accurately predict these cross-domain image data, rendering them insufficient in terms of generalization and practicality for meeting actual production needs [20].

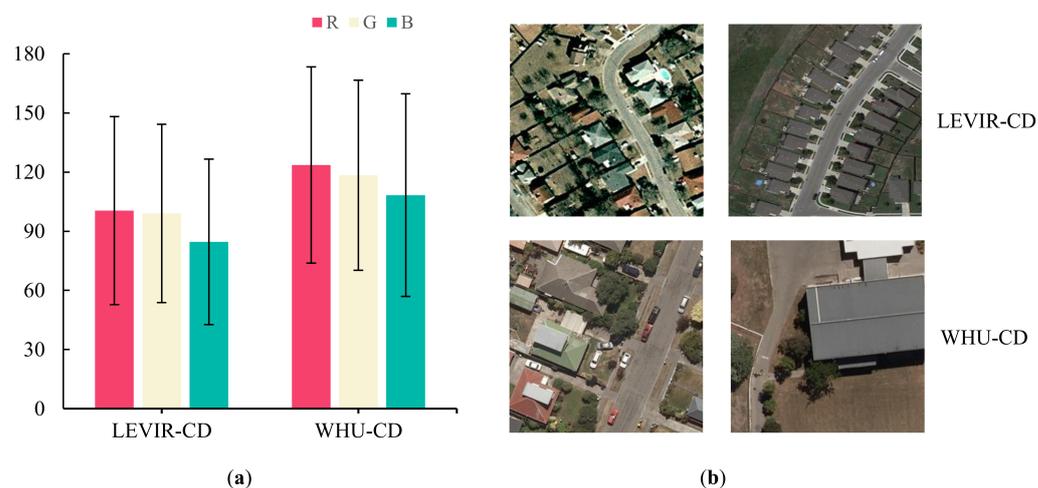


Figure 1. Differences between LEVIR-CD dataset and WHU-CD dataset: (a) is the mean and standard deviation of each channel; (b) is the display of some images of the two datasets.

To enhance the model's generalization capacity, it is advisable to employ specific data augmentation techniques. These may include geometric transformations, such as image translation, horizontal flipping, and rotation, which serve as effective methods for data augmentation. Additionally, spectral data augmentation strategies can be applied, involving adjustments to image brightness, contrast, and the introduction of random noise. However, despite these efforts, specific data augmentation methods do not entirely

bridge the appearance disparities across diverse datasets. In the actual prediction process, relying solely on data augmentation to enhance the model's generalization capability reveals inherent limitations. It is unrealistic to anticipate that data augmentation alone can sufficiently align the transformed training data with the distribution of all testing data. Effectively addressing the substantial diversity in appearance necessitates the adept design of deep learning frameworks. This stands as a pivotal challenge in the realm of cross-domain change detection.

In this paper, we present an alternative approach to tackle the challenge of cross-domain image change detection and enhance the model's generalization capabilities. Our methodology incorporates the use of instance normalization (IN) [21], a neural network module that adeptly captures variations in image appearances within the change detection network while preserving discriminative attributes. Notably, our change detection model strategically combines IN with batch normalization (BN) [22] as fundamental building blocks, augmenting its learning and generalization abilities [23].

Our model, denoted as CrossCDNet, demonstrates notable advantages in cross-domain detection owing to its robust generalization capability. Unlike conventional change detection models, CrossCDNet extensively learns both spatial and semantic features from images captured before and after the change detection interval, seamlessly unifying them. Subsequent research has further validated the effectiveness of CrossCDNet by leveraging diverse architectural configurations to enhance model accuracy and generalization capabilities.

To further address the issue of inadequate model generalization caused by differences in color domains, we have introduced a global-module-based attention mechanism into the encoding–decoding structure of the neural network. This mechanism is designed to compensate for the encoder module's tendency to focus only on central features by introducing an attention mechanism that takes into account the global context. The experimental results demonstrate that this mechanism can aggregate and refine semantic-level features, thereby improving the model's performance.

The contributions of this paper are mainly as follows:

- (1) We propose a Siamese network named CrossCDNet for remote sensing change detection, which is based on IN and BN of IBNM. CrossCDNet significantly enhances the cross-domain detection capability of the change detection model.
- (2) This paper employs a global attention mechanism to address the shortcomings of IN, which tends to focus more on central features.
- (3) The experimental results indicate that CrossCDNet exhibits competitive performance in change detection tests compared to mainstream change detection models and has better generalization ability in cross-evaluation. The code for CrossCDNet will be open-sourced at <https://github.com/XJCXJ/CrossCDNet> (accessed on 16 December 2023).

This paper is structured into multiple sections. In Section 2, we present the cross-domain change detection method proposed in this study, accompanied by a description of the dataset and experimental particulars. Section 3 provides a series of quantitative analysis results derived from two comparative experiments. In Section 4, we delve into the significance of both shallow spatial and deep semantic features in the context of cross-domain change detection. Lastly, Section 5 serves as the conclusion for this paper.

2. Methods and Materials

2.1. CrossCDNet

2.1.1. Network Architecture

CrossCDNet comprises four primary components: an encoder, attention module, connection module, and decoder. The foundational structure adheres to the widely embraced encoder–decoder architecture prevalent in contemporary change detection models. Specifically, the input segment of the network structure conforms to the standard Siamese neural network architecture. In this configuration, we separately input the two temporal images (T1 and T2) into the encoders of the Siamese network's two branches. Throughout each downsampling

step, uniform convolutional kernels are employed for feature extraction from the feature maps within both branches, activating corresponding positions within these feature maps.

Figure 2 illustrates the use of identical encoders for the preservation of both spatial and semantic information of the two branches during the four-stage downsampling process of the paired temporal images. Afterward, in the “Fusion” layer, we connect T1 (preperiod image) and T2 (posterior image) by using a “Concatenation” technique. To be more specific, we perform calculations by pairing the feature maps of the two branches of T1 and T2 based on their hierarchical structure:

$$OUT = Cat[F_A, F_B] \quad (1)$$

where *Cat* module in PyTorch is a connection method that does not alter the length and width of an image but rather only adds the number of channels of the two tensors together. F_A and F_B represent the feature maps. The fused results are then fed into a simple decoder composed of an MLP. During the model training phase, we continually refine our model by minimizing both the distance map and the ground truth label map. This process ensures that distance values for change points become larger, while those for unchanged points decrease. In the testing phase, we obtain the predicted label map, denoted as P , through straightforward threshold calculations applied to the distance map.

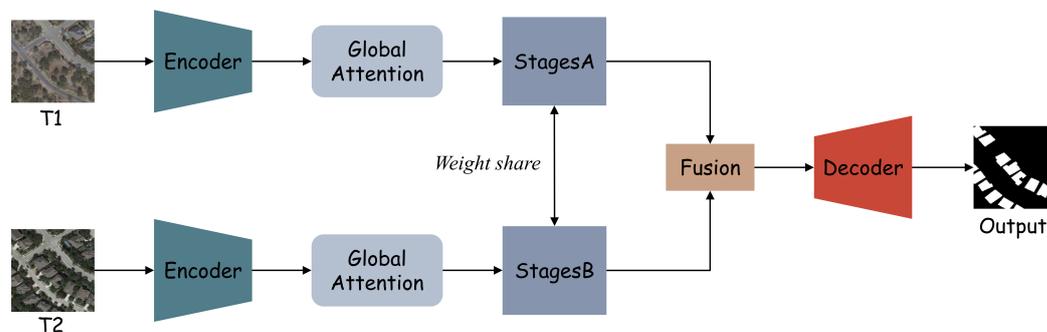


Figure 2. CrossCDNet model structure.

2.1.2. Encoder

In previous research, CNNs have been proposed to enhance the learning of spatial and semantic features. Due to their robust performance in computer vision, CNN-based methodologies have found extensive applications in remote sensing [24]. The domain of change detection, in particular, has seen the emergence of several outstanding encoder networks designed for feature extraction. For instance, TINYCD [25], a compact and lightweight feature extraction network, prioritizes the significance of low-level features in image analysis while maintaining a minimal network parameter count. Another lightweight model for change detection, LightCDNet [11], adopts ShuffleNet v2 as its lightweight backbone network for feature extraction, complemented by a pyramid decoder for end-to-end change detection. The Feature Interaction Network model, known as Changer [26], introduces a groundbreaking universal change detection architecture named MetaChanger, incorporating a series of alternative interaction layers within the feature extractor. Furthermore, the Hierarchical Attention Network (HANet) has been developed [27], which integrates multiscale features and refines detailed features.

Our work draws inspiration from these networks in constructing a feature extractor. We have made enhancements to the encoder’s structure based on ResNet. However, given the requirement of our change detection task to accommodate datasets spanning different color domains, we employ a normalization method distinct from those used in the networks. Within our encoder, we employ a module that combines both BN and IN.

BN is a technique that mitigates the issue of “internal covariate shift” during neural network training by integrating normalization into the model structure. This technique facilitates larger learning rates and faster training speeds. During training, BN standardizes

each feature channel by calculating the mean and variance of a small batch of data. During inference, it applies global statistics to normalize features. BN has become a standard component in the most popular CNN architectures, including ResNet.

Unlike BN, IN normalizes the spatial and semantic features of images using statistics from a single sample instead of a small batch of samples. Additionally, IN applies the same normalization process during both training and inference, enabling the filtering of instance-specific contrast information from the content.

$$\text{IN}(\mathbf{F}) = \frac{\mathbf{F}_{n,k} - \mu_{n,k}}{\sigma_{n,k} + \varepsilon} \quad (2)$$

where $\text{IN}(\cdot)$ denotes the instance normalization process and ε is a small value to avoid division by zero. Then, mean $\mu_{n,k}$ and standard deviation $\sigma_{n,k}$ of n -th sample k -th channel are computed as follows:

$$\mu_{n,k} = \frac{1}{HW} \sum_{h=1}^H \sum_{w=1}^W \mathbf{F}_{n,k,h,w} \quad (3)$$

$$\sigma_{n,k} = \sqrt{\frac{1}{HW} \sum_{h=1}^H \sum_{w=1}^W (\mathbf{F}_{n,k,h,w} - \mu_{n,k})^2} \quad (4)$$

where H is the height and W is the width in Equations (3) and (4).

Firstly, through the concurrent application of IN and BN, our model acquires the statistical benefits associated with feature divergence across various network depths. Secondly, IN ensures visual appearance invariance, while BN expedites training and safeguards discriminative features. The amalgamation of these advantages plays a pivotal role in our model architecture by augmenting its learning and generalization capabilities, all while preserving computational efficiency. Leveraging IN empowers our model to adeptly capture variations in the visual appearance of images within the network, concurrently preserving discriminative attributes and thus amplifying the model's learning and generalization capacities.

CrossCDNet is a new network based on the ResNet model that incorporates different normalization techniques and has been improved from the original ResNet model. The schematic diagram of the CrossCDNet backbone structure as an encoder is shown in Figure 3a, while Figure 3b illustrates the convolutional processing module of the initial images in two periods. We used IN for the initial convolutional processing, and the processed result was input into the residual block. Additionally, we designed an IN and BN module (IBNM) to optimize the image's feature structure, as depicted in Figure 3c. The first two layers of feature maps' spatial feature information is considered more significant in the network architecture. Therefore, we used IBNM for the initial two layers of the backbone while retaining the original BN for the deeper third and fourth layers. This strategy improves the performance of our network in predicting cross-domain images during the inference stage.

Our design aims to enhance the performance of the model in recognizing cross-domain images and generalizing well. Using IN has a two-fold benefit. First, it allows the model to learn invariant features of shallow appearances, improving its ability to use images with high appearance diversity. Second, the moderate incorporation of IN at appropriate positions in the network structure further enhances the model's ability to recognize cross-domain images and generalize well.

2.1.3. Attention Module

Although IBNM/IN achieves center alignment of the feature distribution, it fails to align the global distribution of the features. Therefore, we introduced a global attention mechanism at the end of the encoder to enable the model to better leverage global spatial information.

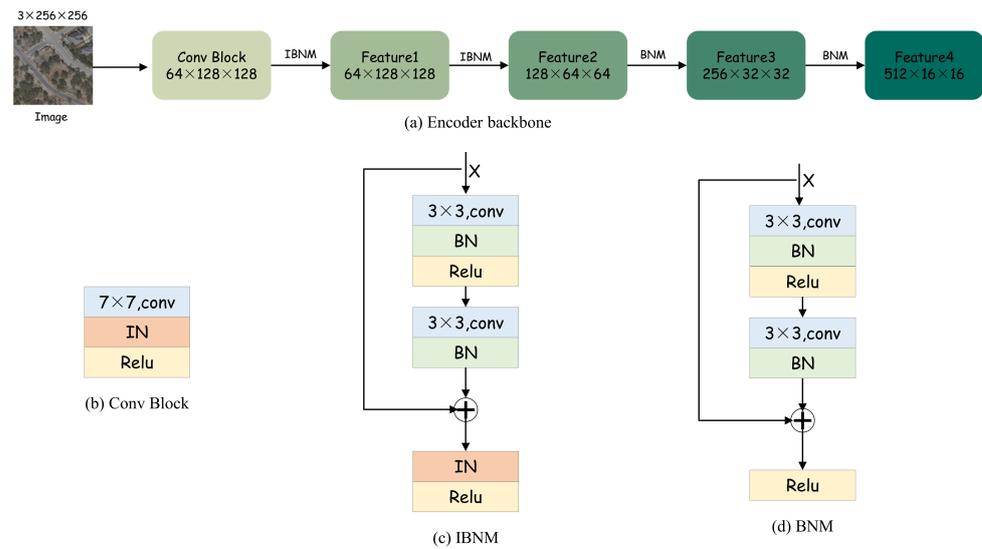


Figure 3. CrossCDNet encoder structure: (a) is encoder backbone; (b–d) are the modules in the backbone.

As illustrated in Figure 4, C , H , W represent the channel count, length, and width of the feature map, respectively. The encoder produces feature maps comprising four layers. In Figure 2, we have performed a reinforcement operation using this attention mechanism for the feature maps generated at each layer of the encoder. Consequently, the feature maps fed into the attention network exhibit four distinct shapes: $(64, 128, 128)$, $(128, 64, 64)$, $(256, 32, 32)$, and $(512, 16, 16)$. ‘Conv’ denotes `torch.nn.conv2d()`, with the following numbers indicating the quantity of convolutional kernels. ‘LN’ signifies the implementation of a normalization layer utilizing the Layer Norm type. Post LN module construction, its resulting shape aligns with the input shape. ‘Parameter’ refers to `torch.nn.Parameter`, a subclass of `torch.Tensor` utilized primarily as trainable parameters in `nn.Module`. A critical distinction from `torch.Tensor` lies in `nn.Parameter` automatic recognition as a trainable parameter in `nn.Module`, i.e., its inclusion in the iterator parameter(). In contrast, regular tensors within the module, not identified as `nn.Parameter()`, are excluded from the parameters. Notably, we quote this attention mechanism from <https://github.com/likyoo/open-cd/blob/main/opencd/models/backbones/tinytinet.py> (accessed on 16th December 2023).

2.1.4. Decoder

In this paper, we use a lightweight decoder called SegformerHead, composed only of MLP layers. The specific structure of this decoder is depicted in Figure 5. Initially, multilevel features (referred to as “Features”) extracted from the encoder and connection layers are processed through MLP layers, consolidating the channel dimensions. Then, in the second step, each feature map is unsampled to one-fourth of the original size (with one-fourth of the original number of channels), followed by concatenating all feature maps. Finally, another MLP layer leverages these fused features to predict classification masks with an $H \times W \times N_{cls}$ resolution, where N_{cls} denotes the number of classes. In the context of change detection tasks, N_{cls} is typically set to 2, discerning between change and no change. The SegformerHead structure employed in our work is characterized by simplicity, composed exclusively of MLP layers, thereby alleviating the need for manual design, and reducing computational demands.

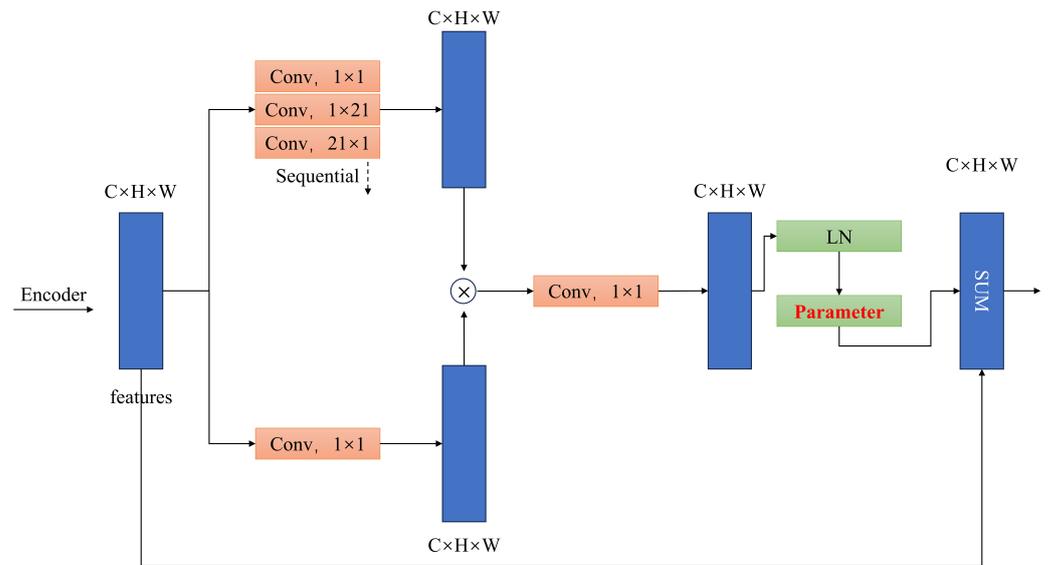


Figure 4. Global attention mechanisms.

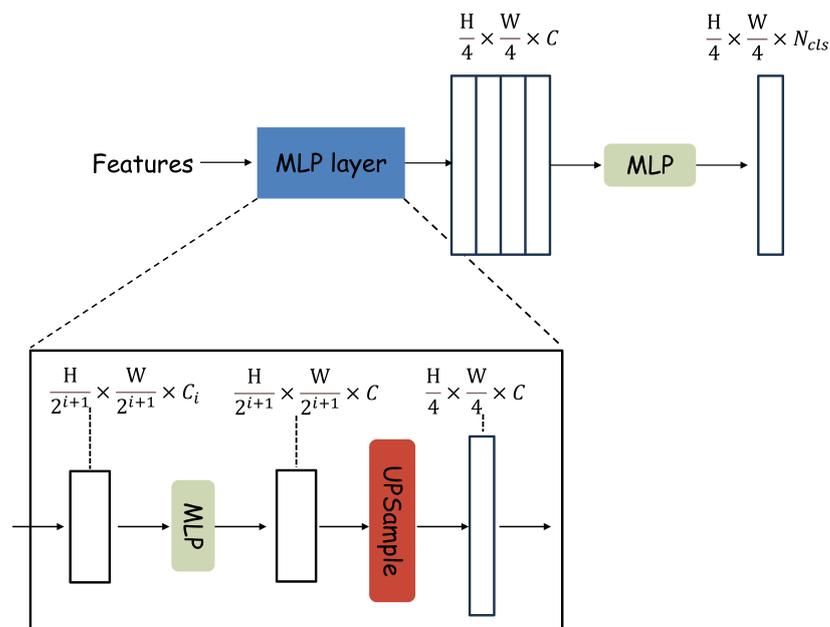


Figure 5. Detailed structure of the SegformerHead decoder.

2.1.5. Details of Loss Function

Imbalance between positive and negative samples is often observed in change detection datasets. For instance, in the popular change detection dataset LEVIR-CD [10], positive samples account for merely 4.65%, while the proportion of positive samples is as low as 4.33% in WHU-CD [28]. This phenomenon is not exclusive to building change detection datasets; it is also present in forest change detection datasets, where an imbalance between positive and negative samples is observed. Specifically, the pixel area of the changed regions is significantly smaller than that of the unchanged regions. To mitigate the impact of sample imbalance and improve model performance in change detection, we employ the strategy of online hard example mining (OHEM) during network training with a cross-entropy loss function:

$$L_{CE}(p_i, q_i) = - \sum_{i=1}^c p_i \log(q_i) \tag{5}$$

$$L_{oss}(p_i, q_i) = \text{OHEM}(L_{CE}) \quad (6)$$

where C is used to denote the number of categories, p_i to represent the ground truth, and q_i to signify the model prediction. To enhance the effectiveness of network parameters, OHEM can be explained in the following three steps:

- (1) The OHEM algorithm selects hard examples that are difficult to train as training samples. Hard examples refer to samples with diversity and high losses.
- (2) The selection of hard examples is based on the loss value of each ROI, where those with the highest loss are chosen. In practical implementation, the initial single ROI Network is augmented into two ROI Networks, with shared parameters. The first ROI Network exclusively executes forward operations, primarily serving the purpose of loss computation. The second ROI Network incorporates both forward and backward operations. It takes challenging examples as input, computes loss, and conducts gradient backpropagation.
- (3) Additionally, the losses between ROIs with high intersection-over-union (IoU) values are relatively similar. As a result, we use nonmaximum suppression (NMS) to eliminate ROIs with a high IoU. The threshold for sample screening is set to 0.7.

This algorithm's strength resides in its ability to mitigate imbalanced data categories without necessitating the adjustment of positive and negative sample ratios. Additionally, with the expansion of the dataset, the algorithm's enhancement becomes more conspicuous [29].

2.2. Dataset and Evaluation Metrics

In this section, we will introduce the experimental data and the research environment, conduct model comparisons, and evaluate metrics, and subsequently engage in a detailed discussion on the experimental outcomes pertaining to individual datasets as well as cross-domain datasets. It should be noted that the cross-domain datasets used in this context were not included in the training process.

2.2.1. Dataset

We conducted a series of comparative experiments to verify the effectiveness of our method using the commonly used public evaluation datasets, LEVIR-CD and WHU-CD, in the field of change detection. Our model was trained on the LEVIR-CD dataset and evaluated on both the LEVIR-CD and cross-domain WHU-CD test sets. The LEVIR-CD dataset comprises 637 pairs of high-resolution images, each with a pixel density of 0.5 m. Each image measures 1024×1024 pixels. These images were systematically gathered from 20 distinct regions across various cities in the United States, encompassing Austin, Lakeway, Bee Cave, Buda, Kyle, Manor, Pflugerville, Dripping Springs, and several other locations within the state of Texas. The dataset was partitioned regarding the official method into training, validation, and test sets, with a ratio of 7:1:2. The WHU-CD dataset contains a pair of very high-resolution (0.2 m/px) image patches measuring $32,507 \times 15,345$ px, captured from aerial image data in New Zealand. We used the official data segmentation method to partition the test set, consisting of 744 pairs of 256×256 -px images. Normalized mean values were calculated for each channel of all images in the two datasets, and the results are shown in Figure 1a. Notably, WHU-CD and LEVIR-CD differ significantly in color space, which is an important consideration for studying cross-domain models.

2.2.2. Evaluation Metrics

Accuracy metrics are computed at the pixel level. Hence, to quantitatively assess the CrossCDNet model's performance, we utilized four common metrics: precision, recall, F1-score [30], and IoU, in a comparative experimental study on change detection. Among these metrics, the F1-score, as the harmonic mean of precision and recall, offers a more precise representation of the model's performance on imbalanced datasets. IoU

represents the intersection over union between the detection results and the actual ground truth. The specific calculation formulas are as follows:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (7)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (8)$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

$$\text{IoU} = \frac{TP}{TP+FN+FP} \quad (10)$$

In these equations, TP is the area of pixels where positive samples are correctly predicted, FP is the area of pixels incorrectly predicted as positive samples, and FN is the area of pixels where positive samples are missed.

2.3. Training Set and Implementation Details

We implemented the CrossCDNet model using the PyTorch framework and verified its functionality on Open-CD, a toolbox for change detection. During training, we set the batch size to 8 and used AdamW as the optimizer with a learning rate of 1×10^3 . To further enhance the model's generalization ability, we incorporated simple data augmentation methods, such as random flipping and rotation, during the model training phase. We conducted comparative experiments on the Tesla V100S GPU (32GB), training the model for 40k iterations until it reached convergence. Table 1 shows a comparison of model sizes, and, although CrossCDNet has a larger number of parameters, the above servers we used were still able to accomplish the training and testing tasks quickly.

Table 1. Model size comparison.

Model	Flops (G)	Params (M)
FC-EF	3.244	1.353
FC-SIAM-CONC	4.989	1.548
FC-SIAM-DIFF	4.385	1.352
BIT	8.749	2.99
Changeformer	2.455	3.847
SNUNet	46.697	12.035
Hanet	20.822	3.028
CrossCDNet	20.371	12.569

2.4. Comparison and Analysis

2.4.1. Comparison with Other Models

In this study, we compared several representative change detection methods. FC-EF, FC-Siam-conc, and FC-Siam-diff [31] are UNet-based change detection baseline models that use a straightforward fusion of Siamese neural networks. STANet is a Siamese network that integrates spatiotemporal attention, with the objective of exploring the utilization of spatiotemporal relationships in change detection. SNUNet is a high-performance change detection model that uses dense skip connections to alleviate the problem of location information loss in deep neural networks. BIT [17] is a hybrid model of CNN and transformer that models the context in the spatiotemporal domain efficiently by employing a bitemporal image transformer. ChangeFormer [32] is a conjoined network based on transformers that integrates a hierarchical transformer encoder and an MLP decoder in the conjoined network to effectively present long-distance details. Transformer structures are employed in both BIT and ChangeFormer.

A model demonstrating high test accuracy on untrained datasets holds substantial practical value. The CrossCDNet model was trained using the LEVIR-CD dataset and subsequently validated on the test sets of both LEVIR-CD and WHU-CD.

2.4.2. Compare with Different Structures of CrossCDNet

In a similar vein, to assess the effectiveness of the IBN module proposed in this paper and the utilization of IN, while also expanding the potential applications of CrossCDNet, we conducted specific ablation experiments. These experiments involved modifying the structure of CrossCDNet based on the trials. Within these experiments, CrossCDNet-a signifies the utilization of the IBN module for all downsampling operations in the encoder section, CrossCDNet-b signifies the absence of IBN modules in the encoder part, and CrossCDNet-c indicates the application of BN to the Conv Block depicted in Figure 3a, with the rest of the structure remaining unaltered. Detailed modifications are illustrated in Figure 6.

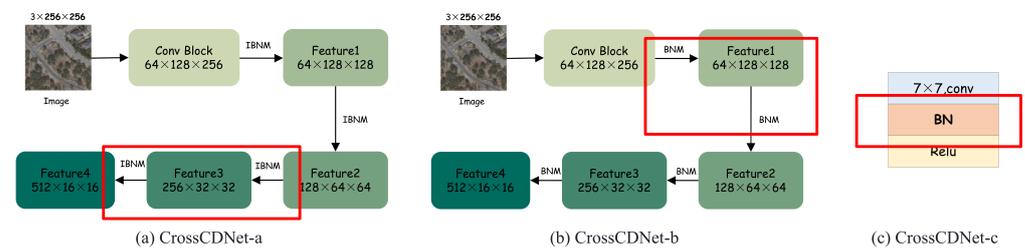


Figure 6. Schematic representation of the structural changes in CrossCDNet. The red boxes show how the models differ from CrossCDNet.

3. Results

3.1. Results of All Methods

In the initial phase, we conducted a comparative analysis to assess the accuracy of various change detection models employing the LEVIR-CD test dataset. This analysis involved calculating precision, recall, F1-score, and IoU concerning detected changes, as elaborated in the introductory section of Table 2. Most of the indicators consistently illustrate that CrossCDNet showcases robust performance and competitiveness. CrossCDNet attains a precision rate of 93.35%, a recall rate of 90.08%, an F1-score of 91.69%, and an IoU of 84.65% across these metrics; all the metrics display optimal performance.

Subsequently, we conducted experiments using a cross-domain dataset. A model demonstrating high test accuracy on untrained data holds significant practical application value. We subjected our model to testing using the WHU-CD dataset, and the test results are presented in the latter part of Table 2. Within the WHU-CD dataset, CrossCDNet achieved precision, recall, F1-score, and IoU values of 77.43%, 76.76%, 77.09%, and 62.73%, respectively. Importantly, our model outperformed all the evaluation metrics in these results. This unequivocally illustrates the superiority of our model in cross-domain detection and its exceptional generalization capabilities.

In the realm of cross-domain detection, the CrossCDNet model demonstrates superior precision concerning the F1-score and IoU metrics. Thus, this experimental outcome substantiates the efficacy of the two methodologies employed in this study for enhancing the model's capacity for generalization. This enhancement has the potential to elevate the performance of change detection models in the context of cross-domain detection.

Table 2. Comparison table of ablation experiment results (training set: LEVIR-CD).

Method	LEVIR-CD (%)				WHU-CD (%)			
	Precision	Recall	F1-Score	IoU	Precision	Recall	F1-Score	IoU
FC-EF [31]	86.91	80.17	83.40	71.53	19.70	62.52	29.96	17.26
FC-Siam-diff [31]	89.53	83.31	86.31	75.92	41.51	51.41	45.53	29.47
FC-Siam-conc [31]	91.99	76.77	83.69	70.96	49.59	44.29	46.79	30.54
BIT [17]	89.18	87.17	88.16	78.83	73.04	66.00	53.07	53.07
ChangeFormer [32]	89.24	89.37	89.31	80.68	65.73	53.41	58.93	41.78
SNUNet [16]	92.11	90.07	91.08	83.61	58.41	71.16	64.15	47.23
HANet [27]	91.21	89.36	90.28	82.27	57.50	67.65	62.16	45.10
CrossCDNet	93.35	90.08	91.69	84.65	77.43	76.76	77.09	62.73

Color convention: **best**, **2nd-best**, and **3rd-best** for all models.

In order to verify the effectiveness of our method, we conducted a reverse experiment to test the portability of the model. Therefore, we conducted the same experiments described above using WHU-CD as the training dataset and LEVIR-CD as the test dataset. The results of the test are shown in Table 3. CrossCDNet also shows strong competitiveness regarding the model trained by WHU-CD. When WHU-CD is the test set, the indexes are precision 95.79%, recall 91.96%, F1-score 93.83%, and IoU 88.38%, and the comprehensive evaluation index F1-score is only 0.15% lower than the first ChangerFormer. Meanwhile, in the cross-domain dataset LEVIR-CD test, CrossCDNet ranked third in accuracy rate with 79.24%, and the rest of the indicators are all optimal. Combined with the detection results in Table 1, CrossCDNet has excellent modeling ability, detection accuracy, and model generalization ability.

Table 3. Comparison table of results of ablation experiments (training set: WHU-CD).

Method	WHU-CD (%)				LEVIR-CD (%)			
	Precision	Recall	F1-Score	IoU	Precision	Recall	F1-Score	IoU
FC-EF	81.69	68.97	74.80	59.74	73.52	4.53	8.54	4.46
FC-Siam-diff	46.24	74.59	57.09	39.95	86.35	4.47	8.51	4.44
FC-Siam-conc	39.90	85.77	54.46	37.4	67.73	4.80	8.96	4.69
BIT	87.45	91.88	89.61	81.18	64.41	10.44	17.97	9.87
ChangeFormer	96.42	91.66	93.98	88.64	86.61	39.35	54.11	37.09
SNUNet	88.72	86.01	87.34	77.53	33.71	6.73	11.22	5.94
HANet	89.07	87.72	88.39	79.20	21.17	7.89	11.50	6.10
CrossCDNet	95.79	91.96	93.83	88.38	79.24	41.73	54.67	37.62

Color convention: **best**, **2nd-best**, and **3rd-best** for all models.

The visualization results pertaining to change detection are presented in Figures 7 and 8. Figure 7 showcases the visualization outcomes of various models applied to the LEVIR-CD dataset, while Figure 8 illustrates the visualization outcomes for diverse models on the WHU-CD dataset. Upon scrutinizing the visual outputs from both datasets, it becomes evident that the detection results obtained by CrossCDNet exhibit a closer alignment with the ground truth labels when compared to alternative models. This superiority is manifested through a reduction in the occurrence of false positive and false negative regions. It is noteworthy that this outcome corresponds to the precision and recall values reported in Table 2.

The CrossCDNet model distinguishes itself from other change detection models by leveraging the statistical benefits of feature divergence at various depths through the simultaneous utilization of IN and BN. This strategy upholds visual appearance invariance while preserving discriminative features in images. The amalgamation of these techniques elevates the model's capacity for learning and generalization, all the while keeping the computational costs constant. The quantified outcomes presented in Tables 1 and 2, along with the visual representations in Figures 7 and 8, precisely affirm this distinctive feature.

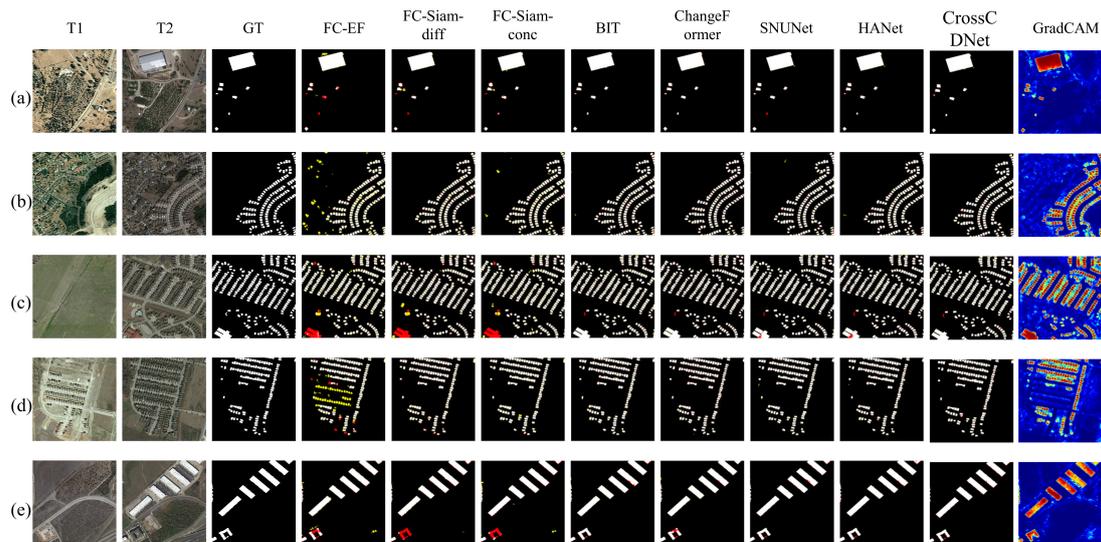


Figure 7. A visual comparison of various methods is conducted on the LEVIR-CD test dataset. To enhance clarity and convenience, different colors are utilized to facilitate a more distinct visualization of the results: true positives (TP) are represented in white, false positives (FP) in yellow, true negatives (TN) in black, and false negatives (FN) in red. Subfigures (a–e) are five sets of images selected from the LEVIR-CD.

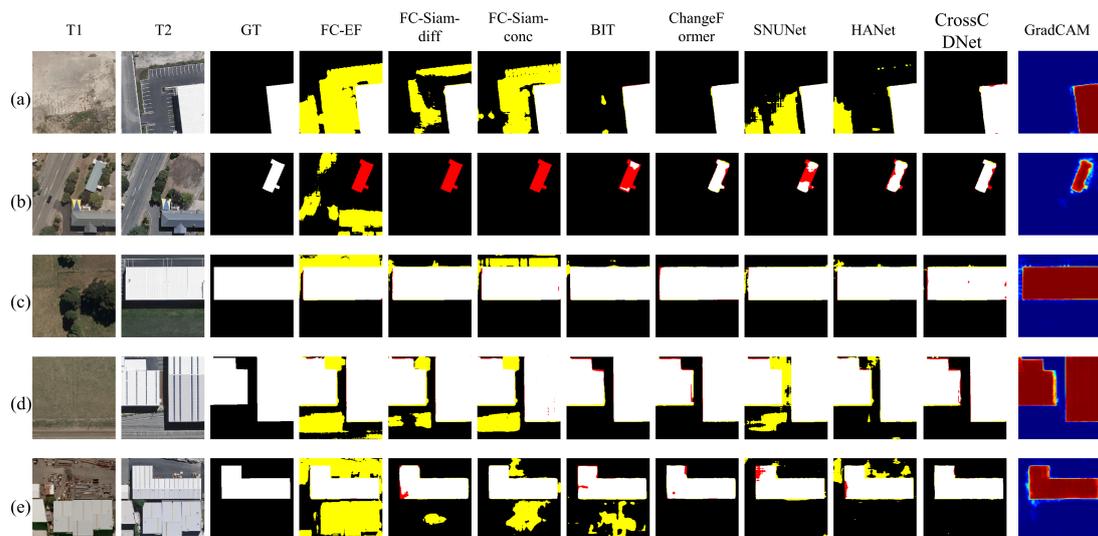


Figure 8. A visual comparison of various methods is conducted on the WHU-CD test dataset. True positives (TP) are represented in white, false positives (FP) in yellow, true negatives (TN) in black, and false negatives (FN) in red. Subfigures (a–e) are five sets of images selected from the WHU-CD.

3.2. Results of Ablation Experiments

We leverage IBNM in a versatile manner to craft diverse architectural configurations for CrossCDNet, thereby enhancing its adaptability in numerous contexts. Following the experimental approach detailed in Section 3.1, we conducted a comprehensive performance evaluation of our model using both the LEVIR-CD and WHU-CD datasets. The specific comparative outcomes are documented in Table 4. Among the tested models, the CrossCDNet-a architecture exhibits superior performance on the LEVIR-CD dataset, achieving the highest recall, F1-score, and IOU. Only the F1-score and IOU values of the CrossCDNet-a architecture surpass CrossCDNet. The results obtained from the cross-domain dataset underscore the advantages of the CrossCDNet architecture, with all four metrics excelling, except for

precision. This ablation experiment thoroughly validates the superior performance of CrossCDNet in cross-domain change detection and the adaptability of its variant architectures in diverse environmental contexts.

Table 4. Comparison table of ablation experiment results.

Method	Precision	LEVIR-CD (%)			WHU-CD (%)			
		Recall	F1-Score	IoU	Precision	Recall	F1-Score	IoU
CrossCDNet-a	92.70	90.73	91.70	84.68	78.58	74.63	76.56	62.02
CrossCDNet-b	92.76	90.37	91.55	84.41	76.95	70.54	73.61	58.24
CrossCDNet-c	93.34	90.04	91.66	84.60	71.15	72.61	71.87	56.10
CrossCDNet	93.35	90.08	91.69	84.65	77.43	76.76	77.09	62.73

Color convention: **best** and **2nd-best**.

Figures 9 and 10 depict the visualized results of various CrossCDNet models applied to the LEVIR-CD and WHU-CD datasets, respectively. Notably, Figure 10 reveals a close alignment between the CrossCDNet model's cross-domain change detection outcomes and the ground truth label, particularly in the context of the WHU-CD test dataset. However, when examining the detection results of the CrossCDNet-c architecture, Figure 10e conspicuously exhibits a significant number of false positives. This issue primarily stems from the absence of the IN module within the Conv Block connection block, as illustrated in Figure 5c.

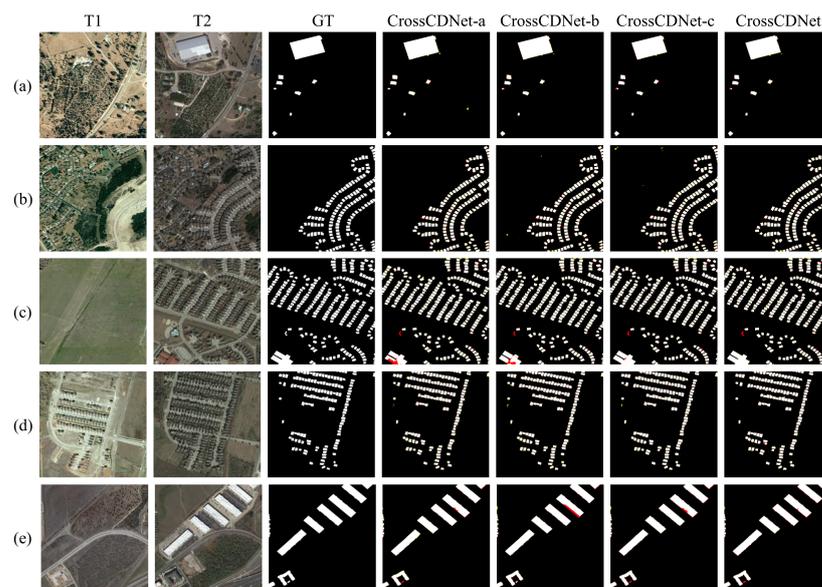


Figure 9. Visualization results of ablation experiments in the LEVIR-CD dataset. True positives (TP) are represented in white, false positives (FP) in yellow, true negatives (TN) in black, and false negatives (FN) in red. Subfigures (a–e) are five sets of images selected from the LEVIR-CD.

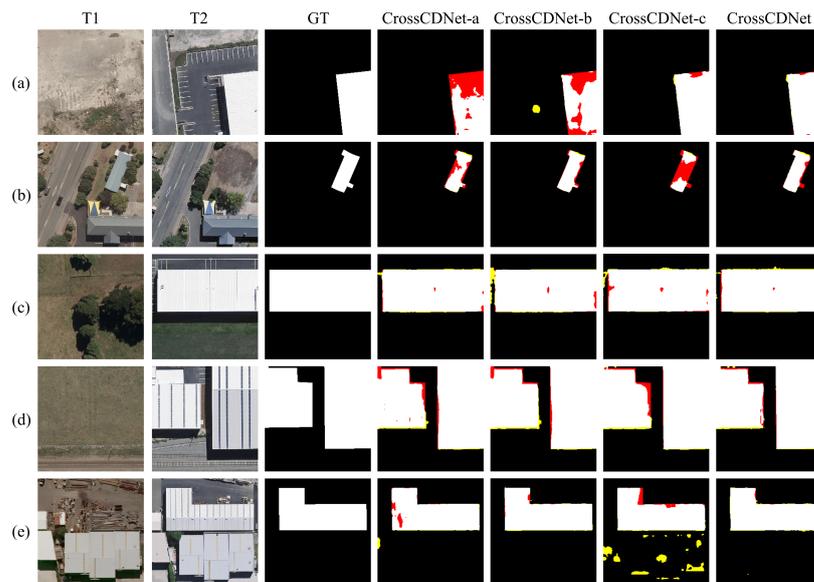


Figure 10. Visualization results of ablation experiments in the WHU-CD dataset. True positives (TP) are represented in white, false positives (FP) in yellow, true negatives (TN) in black, and false negatives (FN) in red. Subfigures (a–e) are five sets of images selected from the WHU-CD.

4. Discussion

In this section, we shall conduct an analysis and discussion regarding the findings presented in Section 3. Our objective is to investigate a fundamental question in the realm of cross-domain change detection: which are more important, shallow spatial features or deep semantic features?

In order to address the issue of comparison, we have introduced a novel architecture named CrossCDNet' as shown in Figure 11, which distinguishes itself from the original CrossCDNet by exclusively employing IBNM in the deeper layers, specifically layers 3 and 4, while abstaining from its use in layers 1 and 2. Below, we will provide a detailed description of its structure and present the results obtained using WHU-CD.

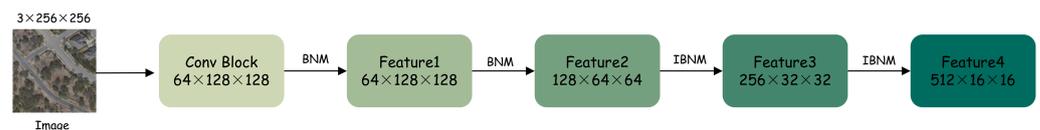


Figure 11. Encoder network structure of CrossCDNet'.

This structure, distinct from the original CrossCDNet, enables a more profound exploration of the IBNM module's influence on either shallow or deep network features.

Upon comparing the outcomes from various networks, the specific detection accuracy and visualization results are presented in the table below. Among all the precision assessment metrics, the CrossCDNet architecture consistently yields the highest values, signifying its superior performance in cross-domain change detection when compared to CrossCDNet'. When scrutinizing the results of CrossCDNet-b in Table 5, it becomes evident that, apart from a slightly elevated precision as compared to CrossCDNet', all the other metrics exhibit lower values. This outcome underscores the notion that the IBNM module is better suited for shallow networks. In the cross-domain change detection network model introduced in this paper, shallow spatial features are deemed more pivotal than deep semantic features, although employing IBNM in deep semantic features does yield certain benefits.

Table 5. Comparison of CrossCDNet with CrossCDNet-b and CrossCDNet' detection results in WHU-CD.

Method	Precision (%)	Recall (%)	F1-Score (%)	IoU(%)
CrossCDNet	77.43	76.76	77.09	62.73
CrossCDNet-b	76.95	70.54	73.61	58.24
CrossCDNet'	75.48	73.15	74.3	59.1

Color convention: **best** and **2nd-best**.

5. Conclusions

In this paper, we propose a Siamese neural network, named CrossCDNet, for cross-domain change detection. We formed the IBNM by combining IN and BN in a specific order and incorporated it as a part of the encoder backbone in the twin neural network to extract feature maps at each layer. In comparison to other models, CrossCDNet exhibits noteworthy advancements in modeling capability and detection accuracy on one dataset while also demonstrating better generalization ability on another dataset. We compare various combinations of IBNM and attention mechanisms utilized by CrossCDNet to address diverse application scenarios.

In the initial phase, we conducted a comprehensive assessment of multiple models using the LEVIR-CD dataset and the WHU-CD dataset, encompassing FC-EF, FC-Siam-diff, FC-Siam-conc, BIT, ChangeFormer, SNUNet, HANet, and CrossCDNet. Following both quantitative and qualitative analysis of diverse accuracy metrics such as precision, recall, F1-score, and IOU, as well as a meticulous review of the visual outcomes for each model, our findings can be summarized as follows: CrossCDNet is an excellent-performing model, exhibiting the closest alignment with ground truth labels on both the LEVIR-CD dataset and the cross-domain detection dataset WHU-CD. On the LEVIR-CD dataset, CrossCDNet achieves a precision of 93.35%, recall of 90.08%, an F1-score of 91.69%, and an IOU of 84.65%. Meanwhile, on the WHU-CD dataset, the corresponding metrics stand at a precision of 77.43%, recall of 76.76%, F1-score of 77.09%, and IOU of 62.73%. This outcome unequivocally underscores the superior modeling capacity, detection precision, and generalization capabilities of CrossCDNet.

Subsequently, we introduced modifications to the encoder component of the CrossCDNet architecture and conducted a series of ablation experiments. As a result of these experiments, it became evident that, within the domain of cross-domain change detection, specifically concerning the WHU-CD dataset, CrossCDNet exhibited the most favorable overall accuracy metrics. Remarkably, in the evaluation of the LEVIR-CD dataset, the structure that outperformed in terms of overall accuracy metrics was recognized as CrossCDNet-a. Simultaneously, while examining the visual results, a high degree of consistency with the quantitative accuracy evaluations was noted.

In the concluding section, we delve into the inquiry, "In the context of cross-domain change detection, which holds greater significance: shallow spatial features or deep semantic features?" Following extensive deliberation and empirical investigations, our findings have unveiled that, within the domain of cross-domain change detection, IBNM proves to be better suited for handling shallow spatial features. Consequently, in the cross-domain change detection network model introduced in this research, it becomes evident that shallow spatial features bear greater importance than their deep semantic counterparts.

Moving forward, we will concentrate on studying the application of the decoder structure in cross-domain change detection and enhancing the generalization capability of our model even more.

Author Contributions: Y.S. supervised the study and revised the manuscript; D.M. supervised the study and provided funding support for the project; J.X. wrote the manuscript and designed the comparative experiments; J.J., W.W. and E.Y. revised the manuscript and provided comments and suggestions for the manuscript; J.J. assisted J.X. in designing the architecture and conducting experiments. All authors have read and agreed to the published version of the manuscript. All authors have read and approved the final version.

Funding: This research was funded by the Project Technology Innovation Plan Project of Hunan Provincial Forestry Department under Grant XLK202108-8, and the Project of Central South Inventory and Planning Institute of State Forestry and Grassland Administration under Grant 68218022 and 2023004, in part by the National Natural Science Foundation of China under Grant 32071682 and Grant 31901311.

Data Availability Statement: The data are available in the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Maxwell, A.E.; Warner, T.A.; Guillén, L.A. Accuracy Assessment in Convolutional Neural Network-Based Deep Learning Remote Sensing Studies—Part 1: Literature Review. *Remote Sens.* **2021**, *13*, 2450. [[CrossRef](#)]
2. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [[CrossRef](#)]
3. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems 28*; Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2015, Volume 28.
4. Zhang, X.; Han, L.; Han, L.; Zhu, L. How Well Do Deep Learning-Based Methods for Land Cover Classification and Object Detection Perform on High Resolution Remote Sensing Imagery? *Remote Sens.* **2020**, *12*, 417. [[CrossRef](#)]
5. Nguyen, N.D.; Do, T.; Ngo, T.D.; Le, D.D. An evaluation of deep learning methods for small object detection. *J. Electr. Comput. Eng.* **2020**, *2020*, 3189691. [[CrossRef](#)]
6. Kussul, N.; Lavreniuk, M.; Skakun, S.; Shelestov, A. Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 778–782. [[CrossRef](#)]
7. Wang, S.; Chen, W.; Xie, S.M.; Azzari, G.; Lobell, D.B. Weakly Supervised Deep Learning for Segmentation of Remote Sensing Imagery. *Remote Sens.* **2020**, *12*, 207. [[CrossRef](#)]
8. Dyson, J.; Mancini, A.; Frontoni, E.; Zingaretti, P. Deep Learning for Soil and Crop Segmentation from Remotely Sensed Data. *Remote Sens.* **2019**, *11*, 1859. [[CrossRef](#)]
9. Wang, P.; Bayram, B.; Sertel, E. A comprehensive review on deep learning based remote sensing image super-resolution methods. *Earth-Sci. Rev.* **2022**, *232*, 104110. [[CrossRef](#)]
10. Chen, H.; Shi, Z. A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection. *Remote Sens.* **2020**, *12*, 1662. [[CrossRef](#)]
11. Xing, Y.; Jiang, J.; Xiang, J.; Yan, E.; Song, Y.; Mo, D. LightCDNet: Lightweight Change Detection Network Based on VHR Images. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 2504105. [[CrossRef](#)]
12. Xiang, J.; Xing, Y.; Wei, W.; Yan, E.; Jiang, J.; Mo, D. Dynamic Detection of Forest Change in Hunan Province Based on Sentinel-2 Images and Deep Learning. *Remote Sens.* **2023**, *15*, 628. [[CrossRef](#)]
13. Jiang, J.; Xiang, J.; Yan, E.; Song, Y.; Mo, D. Forest-CD: Forest Change Detection Network Based on VHR Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [[CrossRef](#)]
14. Lu, P.; Qin, Y.; Li, Z.; Mondini, A.C.; Casagli, N. Landslide mapping from multi-sensor data through improved change detection-based Markov random field. *Remote Sens. Environ.* **2019**, *231*, 111235. [[CrossRef](#)]
15. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Quebec, QC, Canada, 14 September 2017; pp. 3–11.
16. Fang, S.; Li, K.; Shao, J.; Li, Z. SNUNet-CD: A Densely Connected Siamese Network for Change Detection of VHR Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 2504105. [[CrossRef](#)]
17. Chen, H.; Qi, Z.; Shi, Z. Remote Sensing Image Change Detection With Transformers. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5607514. [[CrossRef](#)]
18. Malila, W.A. *Change Vector Analysis: An Approach for Detecting Forest Changes with Landsat*; LARS Symposia, Paper 385; Purdue University Libraries: West Lafayette, IN, USA, 1980.
19. Celik, T. Unsupervised Change Detection in Satellite Images Using Principal Component Analysis and k-Means Clustering. *IEEE Geosci. Remote Sens. Lett.* **2009**, *6*, 772–776. [[CrossRef](#)]
20. Chen, H.; Zhang, H.; Chen, K.; Zhou, C.; Chen, S.; Zou, Z.; Shi, Z. Continuous Cross-Resolution Remote Sensing Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5623320. [[CrossRef](#)]
21. Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Improved Texture Networks: Maximizing Quality and Diversity in Feed-Forward Stylization and Texture Synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
22. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 7–9 July 2015; pp. 448–456.
23. Pan, X.; Luo, P.; Shi, J.; Tang, X. Two at Once: Enhancing Learning and Generalization Capacities via IBN-Net. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.

24. Yuan, Q.; Shen, H.; Li, T.; Li, Z.; Li, S.; Jiang, Y.; Xu, H.; Tan, W.; Yang, Q.; Wang, J.; et al. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sens. Environ.* **2020**, *241*, 111716. [[CrossRef](#)]
25. Codegoni, A.; Lombardi, G.; Ferrari, A. TINYCD: A (not so) deep learning model for change detection. *Neural Comput. Appl.* **2023**, *35*, 8471–8486. [[CrossRef](#)]
26. Fang, S.; Li, K.; Li, Z. Changer: Feature Interaction is What You Need for Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5610111. [[CrossRef](#)]
27. Han, C.; Wu, C.; Guo, H.; Hu, M.; Chen, H. HANet: A Hierarchical Attention Network for Change Detection With Bitemporal Very-High-Resolution Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 3867–3878. [[CrossRef](#)]
28. Ji, S.; Wei, S.; Lu, M. Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 574–586. [[CrossRef](#)]
29. Shrivastava, A.; Gupta, A.; Girshick, R. Training Region-Based Object Detectors With Online Hard Example Mining. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
30. Hand, D.; Christen, P. A note on using the F-measure for evaluating record linkage algorithms. *Stat. Comput.* **2018**, *28*, 539–547. [[CrossRef](#)]
31. Caye Daudt, R.; Le Saux, B.; Boulch, A. Fully Convolutional Siamese Networks for Change Detection. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 4063–4067. [[CrossRef](#)]
32. Bandara, W.G.C.; Patel, V.M. A Transformer-Based Siamese Network for Change Detection. In Proceedings of the IGARSS 2022—2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 207–210. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.