remote sensing

Article

ADF-Net: An Attention-Guided Dual-Branch Fusion Network for Building Change Detection near the Shanghai Metro Line Using Sequences of TerraSAR-X Images

Peng Chen [1,2,3], Jinxin Lin [4,5], Qing Zhao [1,2,3,*], Lei Zhou [1,2,3], Tianliang Yang [4,5], Xinlei Huang [4,5] and Jianzhong Wu [4,5]

[1] Key Laboratory of Geographical Information Science, Ministry of Education, East China Normal University, Shanghai 200062, China; 51213901066@stu.ecnu.edu.cn (P.C.); 51253901073@stu.ecnu.edu.cn (L.Z.)
[2] School of Geographic Sciences, East China Normal University, Shanghai 200241, China
[3] Key Laboratory of Spatial-Temporal Big Data Analysis and Application of Natural Resources in Megacities, Ministry of Natural Resources, Shanghai 200241, China
[4] Key Laboratory of Land Subsidence Monitoring and Prevention, Ministry of Natural Resources, Shanghai 200072, China; ljxsupper@126.com (J.L.); sigs_ytl@163.com (T.Y.); huangxl2009@126.com (X.H.); wjzhongsh@163.com (J.W.)
[5] Shanghai Institute of Geological Survey, Shanghai 200072, China
* Correspondence: qzhao@geo.ecnu.edu.cn; Tel.: +86-21-62224459

Abstract: Building change detection (BCD) plays a vital role in city planning and development, ensuring the timely detection of urban changes near metro lines. Synthetic Aperture Radar (SAR) has the advantage of providing continuous image time series with all-weather and all-time capabilities for earth observation compared with optical remote sensors. Deep learning algorithms have extensively been applied for BCD to realize the automatic detection of building changes. However, existing deep learning-based BCD methods with SAR images suffer limited accuracy due to the speckle noise effect and insufficient feature extraction. In this paper, an attention-guided dual-branch fusion network (ADF-Net) is proposed for urban BCD to address this limitation. Specifically, high-resolution SAR images collected by TerraSAR-X have been utilized to detect building changes near metro line 8 in Shanghai with the ADF-Net model. In particular, a dual-branch structure is employed in ADF-Net to extract heterogeneous features from radiometrically calibrated TerraSAR-X images and log ratio images (i.e., difference images (DIs) in dB scale). In addition, the attention-guided cross-layer addition (ACLA) blocks are used to precisely locate the features of changed areas with the transformer-based attention mechanism, and the global attention mechanism with the residual unit (GAM-RU) blocks is introduced to enhance the representation learning capabilities and solve the problems of gradient fading. The effectiveness of ADF-Net is verified using evaluation metrics. The results demonstrate that ADF-Net generates better building change maps than other methods, including U-Net, FC-EF, SNUNet-CD, A2Net, DMINet, USFFCNet, EATDer, and DRPNet. As a result, some building area changes near metro line 8 in Shanghai have been accurately detected by ADF-Net. Furthermore, the prediction results are consistent with the changes derived from high-resolution optical remote sensing images.

Keywords: building change detection; TerraSAR-X; non-local filtering; improving neighborhood-based ratio; deep learning

Citation: Chen, P.; Lin, J.; Zhao, Q.; Zhou, L.; Yang, T.; Huang, X.; Wu, J. ADF-Net: An Attention-Guided Dual-Branch Fusion Network for Building Change Detection near the Shanghai Metro Line Using Sequences of TerraSAR-X Images. *Remote Sens.* **2024**, *16*, 1070. https://doi.org/10.3390/rs16061070

Academic Editor: Timo Balz

Received: 9 February 2024
Revised: 11 March 2024
Accepted: 15 March 2024
Published: 18 March 2024

1. Introduction

With rapid urbanization, increased urban building changes caused by intensive human activities are noteworthy problems many metropolitan areas face. Monitoring building changes is of great importance for urban planners, as it can determine instability problems of nearby public and private infrastructures. Rapid urbanization has increased underground space utilization to relieve traffic congestion. A growing number of excavation projects passing through the upper part of metro tunnels can quickly destroy the initial stress equilibrium state of the surrounding rock [1–4], especially in Shanghai, where metro tunnels are buried in the soft soil layer [5]. Therefore, timely detection of urban building changes is crucial for monitoring and ensuring ground stability of the surrounding areas of metro lines.

Remote sensing (RS) technologies have widely been used for change detection (CD) using satellite data with various temporal and spatial resolutions. In previous studies, high-resolution optical RS images have mainly been utilized with rich spectral information, clear textures, and regular geometric structures [6–8]. However, most previous studies about automatically monitoring CD using high-resolution optical satellite images were hampered by a) the limited data availability caused by the low temporal resolution and cloud contamination and b) the high spectral variation, complicated imaging conditions, and different viewing angles. These disadvantages have posed significant challenges in CD studies with optical remote sensing images.

In recent years, CD using Synthetic Aperture Radar (SAR) images has become more and more effective for identifying small-dense changes [9–11]. Compared with optical remote sensing images, SAR images have unique advantages for CD. As an active remote sensing system, SAR has all-weather, day–night imaging capability and permits synoptic views of large areas. Meanwhile, a growing number of SAR images acquired by the new generation SAR systems makes it possible to regularly detect building changes in a timely manner. As the new generation of SAR, TerraSAR-X (TSX) has a high spatial resolution and operates with an 11-day revisit time, and thus, it can image building changes with fast dynamics.

The procedure of traditional CD with SAR images is usually divided into the following steps: (1) preprocessing, which is mainly composed of radiometric calibration, multi-looking processing, geocoding, clipping, and conversion to decibels; (2) denoising and difference image (DI) generation; and (3) classification of difference images (DIs) into changed pixels and unchanged pixels. Among these steps, the second step was necessary, affecting the quality of speckle noise suppression. The common denoising methods, including mean filter, median filter, and Gaussian filter, operate on a local area, which makes it challenging to obtain comprehensive receptive field information. Considering that the information extraction is not limited to a local patch, a non-local mean (NLM) algorithm was also proposed, guaranteeing better denoising robustness than local filters [12]. It exploits non-local self-similarity within images to compute the weight of search patches. The classic DI generation methods include the subtraction operator [13] and the log ratio (LR) operator [14,15]. The LR operator is more suitable than the subtraction method since it can transform the multiplicative speckle noise into additive noise. However, these pixel-based methods are poor in suppressing speckle noise. The mean ratio operator [16] uses the mean of local neighborhood pixels to replace its center value to reduce the impact of speckle noise. Gong et al. [17] proposed a neighborhood-based ratio (NR) method for CD with SAR images, which employs heterogeneity of the local area to control the weight of neighborhood information. In the DI classification step, unsupervised and supervised methods are utilized to obtain changed and unchanged regions. In unsupervised methods, thresholding approaches [15] and clustering approaches [18,19] are generally used when the prior information is insufficient. In supervised methods, some traditional techniques take full advantage of previous information to identify changed and unchanged classes, such as support vector machine classification [20], maximum likelihood classification [21], and neural net classification [22].

Deep learning algorithms have recently gained widespread attention in CD, and deep learning models can extract high-level features and effectively reduce the speckle noise of SAR images through the convolution kernels. Long et al. [23] first proposed a fully convolutional network (FCN) for end-to-end semantics segmentation. The FCN used different convolution and pooling layers to capture hierarchical features, and it significantly improved the performance of semantic segmentation tasks. The FCN has been introduced to CD and achieved good results [24,25]. Inspired by the idea of FCN, U-Net was applied to image segmentation and CD as feature extractors [26]. The encoder–decoder structure of U-Net is conducive to multi-scale feature fusion and reducing the cost of computation. In addition, the skip connection can connect fine-grained shallow features with coarse-grained deep features for precise feature learning. Res-UNet, using identity mapping of residual blocks based on U-Net, was proposed to avoid the gradient disappearance and explosion caused by the deepening of the network, and it performs well in urban building change detection (BCD) [9]. Zhou et al. [27] proposed U-Net++, composed of U-Nets of varying depths, and improved the performance with the help of its nested structure and re-design skip connections. Ding et al. [28] proposed a deeply supervised attention-guided network for urban CD. The dual-branch end-to-end network combines multiple different modules to extract more distinctive features. Considering the preservation of edge details, Yang et al. [29] proposed an attention CD network based on Siamese architecture. Its multi-branch structure has the advantage of aggregating similarities, differences, and global information. Basavaraju et al. [30] proposed an encoder–decoder architecture for CD tasks, using modified residual connections and the new spatial pyramid pooling module to fuse multi-level semantic information. A multi-scale capsule network was proposed to extract the discriminative information between unchanged and changed pixels [31], and the designed adaptive fusion convolution works well in weakening the influence of speckle noise. Chen et al. [32] proposed a Stockwell scattering network based on Stockwell transform. It provides noise-resilient feature representation and achieves satisfactory performance on CD. Given the strong neighborhood feature extraction ability of the graph neural network (GNN), Wang et al. [33] proposed an end-to-end dynamic graph-level neural network to utilize the local information of each neighborhood block and learn change features at a graph level for SAR image CD.

Since most previous CD studies with SAR images are limited to simple scenarios or open datasets, which only provide obvious texture information, the existing methods may not be suitable for small-dense building detections with high-resolution SAR image time series of urbanized areas around metro lines. A novel attention-guided dual-branch fusion network (ADF-Net) is proposed for BCD around metro lines. It is applicable for detecting the changes in buildings in complex scenarios, especially around Shanghai metro line 8, connecting the urban and suburban areas. In the proposed framework, we exploit dual-branch architecture to obtain more heterogeneous information using a time series of high-resolution TSX images and DIs on different networks. The swin transformer block [34], global attention mechanism with residual unit (GAM-RU) block, atrous spatial pyramid pooling (ASPP) block [35], and attention-guided cross-layer addition (ACLA) block are introduced into the ADF-Net model for extracting both global and local feature information, then fusing the multi-scale information for detecting building changes more accurately and removing the speckle noise more comprehensively. The contributions of this paper are the following:

- This paper proposes ADF-Net, a novel attentional BCD network based on dual-branch architecture. Parallel networks on both sides excavate change features from SAR intensity images and DIs, respectively.
- Non-local means filtering (NLM) and an improved neighborhood-based ratio (INR) are introduced to generate DIs, effectively reducing the speckle noise of SAR images.
- With the transformer-based attention mechanism, the ACLA block is designed to enhance feature extraction and obtain the precise location of changed areas. The GAM-RU block is proposed to apply the CNN-based attention mechanism to feature maps and simultaneously avoid network degradation. The introduction of ASPP in the auxiliary branch facilitates the extraction of multi-scale features, preserving the detailed information while ignoring the speckle noise.
- The ADF-Net model is applied to detect small-dense building changes in the surrounding areas of metro line 8 in Shanghai.

This paper is organized as follows: Section 2 illustrates the study area and datasets. In Section 3, we introduce the architecture of ADF-Net and a series of blocks used for ADF-Net in detail. The BCD results and comparison with other networks are shown in Section 4. Section 5 presents the BCD results of surrounding areas of metro line 8 in Shanghai. The findings and future works are concluded in Section 6.

2. Study Area and Datasets

2.1. Study Area

Shanghai is located in the east of China, on the southeastern estuary of the Yangtze River Delta. It is one of the megacities in China, with a population of more than 24 million, and has an urban metro system with the longest total length in the world by 2021. There are 20 metro lines in operation with a total length of 831 km. Most metro lines are buried in the soft clay layer, and thus are more vulnerable to ground deformation caused by adjacent buildings or infrastructure constructions. The metro line 8 in Shanghai connects two sides of the Huangpu River. It stretches across the Inner Ring Road and Outer Ring Expressway of Shanghai, linking the urban and suburban areas. The surrounding areas of metro line 8 have been selected as the study area for performing BCD. The location of metro line 8 is shown in Figure 1. It contains twenty-six underground and four elevated stations, totaling 37.5 km from south to north. Furthermore, since metro line 8 is surrounded by different types of buildings (such as urban residences, commercial buildings, etc.), it provides BCD in various complex scenarios to apply ADF-Net.

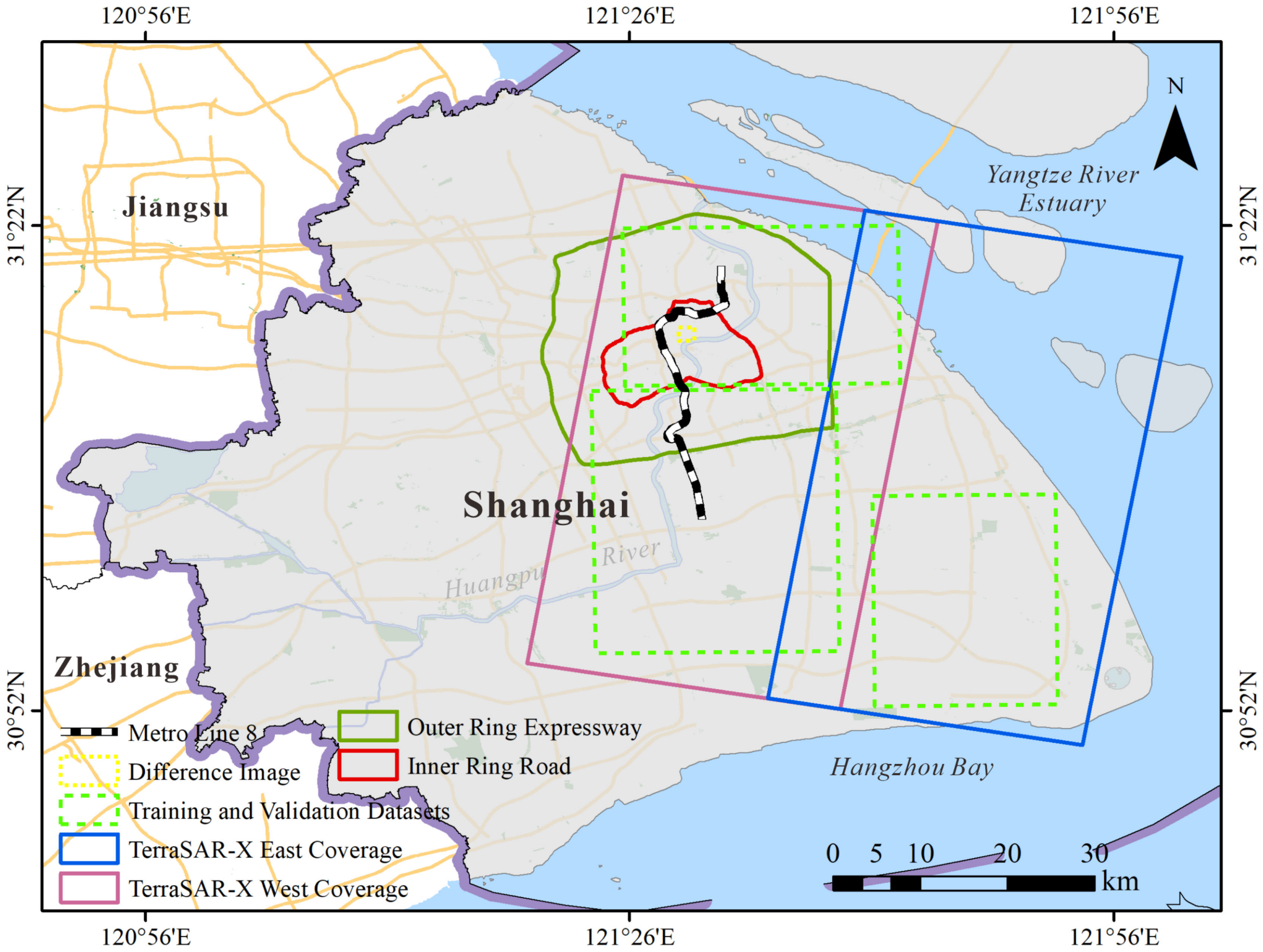

Figure 1. Metro line 8 is located in Shanghai; two descending TerraSAR-X passes are depicted in purple and blue rectangles.

2.2. SAR Images

This study used two descending coverages of twelve TSX images collected in StripMap (SM) mode. Six images were acquired before the changes occurred, and the other six were obtained after the changes. The coverage of TSX images is 50 × 30 km^2, as shown in Figure 1. The resolution of the SM mode image is about 3 m × 3 m (range × azimuth), and the high-resolution TSX images can provide rich details and textures to achieve the detection of small-dense building changes. Multi-looking is conducted to obtain TSX intensity images with 6 m × 6 m resolution to reduce the influence of speckle noise. Other preprocessing includes calibration, co-registration, geocoding, clipping, conversion to decibels, and normalization to 0–255. Table 1 lists the acquisition date of TSX images used in this study.

Table 1. The acquisition dates of TSX images used in BCD.

State	TSX West Coverage	TSX East Coverage
Pre-change	2019-06-19	2015-10-16
	2019-08-24	2015-11-18
	2019-10-29	2015-12-10
Post-change	2021-06-25	2021-06-03
	2021-08-08	2021-07-28
	2021-09-10	2021-09-21

The TSX images are divided into small, non-overlapping patches with a size of 256 × 256 pixels. A total of 493 samples are collected and then augmented to 2958 by performing data augmentation (rotation with 90°, 180° and 270°, and flip vertically and horizontally). Subsequently, the training, validation, and test samples are divided by the ratio of 10:1:1. Thus, 2466 training samples are obtained for our training datasets. The validation datasets include 246 samples for recording the best training epoch and its parameter weights. The test datasets include 246 samples for verifying the BCD accuracy. Each sample includes three pre-change patches and three post-change patches. In addition, the ground truths for training, validation, and test datasets are derived by conducting an improved neighborhood-based ratio on TSX images captured before and after changes, followed by applying a thresholding segmentation method to generate pseudo labels. Subsequently, the quality of pseudo labels is manually enhanced based on Sentinel-2 and GaoFen-7 images.

2.3. Optical Remote Sensing Images

Sentinel-2 and GaoFen-7 images are used to verify the accuracy of BCD around metro line 8. The Sentinel-2 image was acquired on 3 June 2019, with a spatial resolution of 10 m. The GaoFen-7 image was acquired on 3 July 2021, with a spatial resolution of 2.6 m.

3. Methodology

3.1. Basic Architecture of ADF-Net

Unlike other BCD networks, ADF-Net is designed as a dual-branch end-to-end network. The flowchart of the ADF-Net framework is shown in Figure 2. The dual-branch network with the input of heterogeneous information is used to extract more distinctive features. The main advantage of a dual-branch network is that it can learn features of different input images by training two sub-networks simultaneously. In this study, the features of time series TSX images are extracted by the main branch, using encoder–decoder architecture to facilitate learning of the change information. The auxiliary branch extracts features from DIs using a relatively simple network architecture with lower complexity. Through the concatenation operation, the features from different networks are concatenated and fed into the classifier part to enhance the generalization ability of ADF-Net.

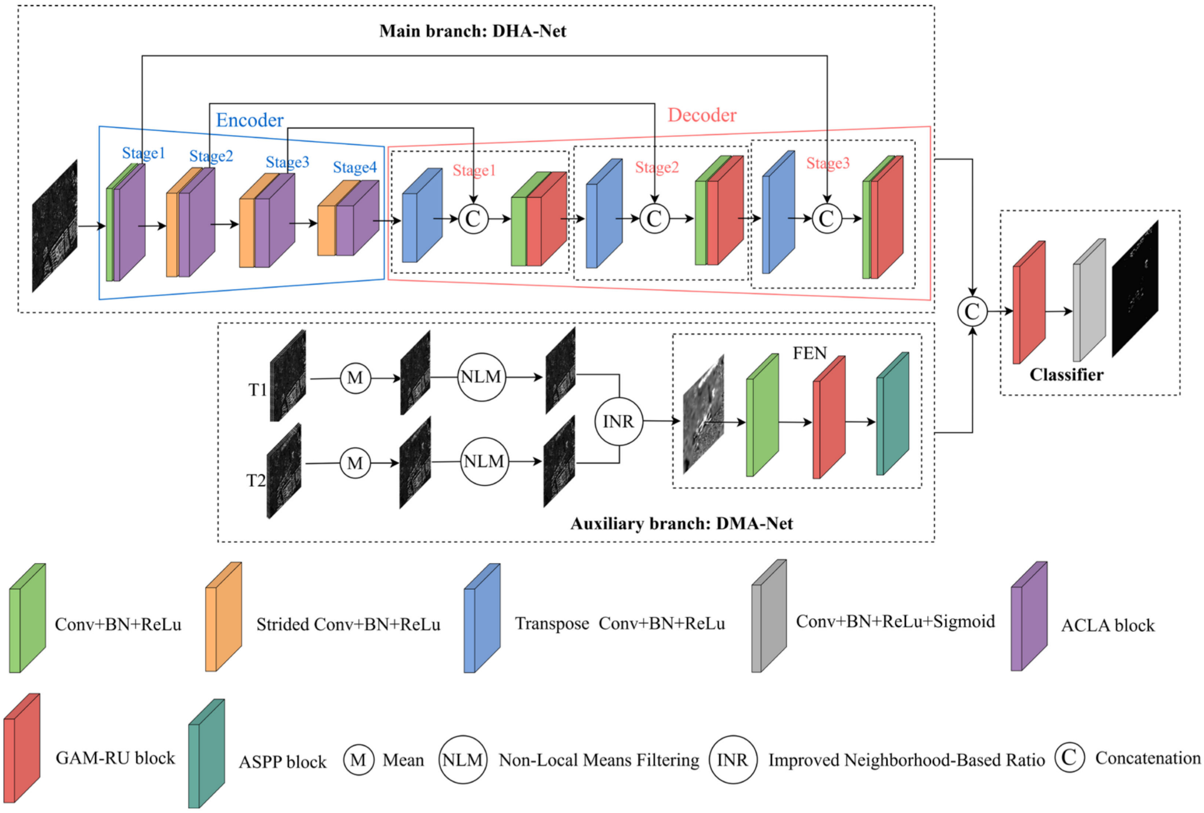

Figure 2. The architecture of ADF-Net for BCD.

Using a deeply hierarchical attentional network (DHA-Net), the main branch extracts multi-scale features from time series TSX images. The structure of DHA-Net is composed of two parts. One is the encoder part. The input is initially entered into one convolutional block consisting of one 3 × 3 convolution, batch normalization, and nonlinearity in the form of ReLu. Then, the output is subjected to three ACLA blocks, each followed by one strided convolutional block to obtain the downsampling of feature maps. At the end of the encoder part, there is one ACLA block for further extracting deep features. The convolutional blocks of U-Net are substituted with ACLA blocks to locate changed areas precisely and assign greater weight to the pixels of changed areas. Strided convolutions are also used instead of pooling for better information retention. Another part is the decoder. The decoder comprises three transpose convolutional blocks, each followed by one convolutional block and one GAM-RU block. The GAM-RU blocks aim to put global attention weights on the input features and remove the problem of vanishing and exploding gradients caused by the deepening of the network. To combine the fine-grained shallow feature maps with the coarse-grained deep feature maps, skip connections between the encoder and the decoder are employed to obtain sufficient information.

The auxiliary branch is introduced to enrich diverse features, using a difference-based multi-scale attention network (DMA-Net) to capture multi-scale difference features. Pixel-wise mean operation is performed on pre-change and post-change time series TSX images, respectively. Given that the speckle noise is an inherent problem of SAR images, their results are processed sequentially by the NLM and INR algorithm to obtain the DIs and improve the quality of DIs by achieving speckle denoising effectively. Then, the DIs are added to a feature extraction network (FEN), which contains one convolutional block, one GAM-RU block, and one ASPP block to extract features and fuse multi-scale features.

In the classifier part, a concatenation operation is applied to the results of DHA-Net and DMA-Net. Then, the concatenated feature maps are input into one GAM-RU block to optimize the network. Finally, a sequence of one 3 × 3 convolution and one sigmoid function is added at the end of ADF-Net to generate change maps. The structure parameters of ADF-Net can be found in Table 2.

Table 2. The structure parameters of ADF-Net.

Stage	Layer	Output Shape
Pre-change images	-	3 × 256 × 256
Post-change images	-	3 × 256 × 256
Encoder-Stage1	Conv ACLA block	36 × 256 × 256 36 × 256 × 256
Encoder-Stage2	Strided Conv ACLA block	72 × 128 × 128 72 × 128 × 128
Encoder-Stage3	Strided Conv ACLA block	144 × 64 × 64 144 × 64 × 64
Encoder-Stage4	Strided Conv ACLA block	288 × 32 × 32 288 × 32 × 32
Decoder-Stage1	Transpose Conv Concatenation Conv GAM-RU block	144 × 64 × 64 288 × 64 × 64 144 × 64 × 64 144 × 64 × 64
Decoder-Stage2	Transpose Conv Concatenation Conv GAM-RU block	72 × 128 × 128 144 × 128 × 128 72 × 128 × 128 72 × 128 × 128
Decoder-Stage3	Transpose Conv Concatenation Conv GAM-RU block	36 × 256 × 256 72 × 256 × 256 36 × 256 × 256 36 × 256 × 256
FEN	Conv GAM-RU block ASPP block	36 × 256 × 256 36 × 256 × 256 36 × 256 × 256
Classifier	Concatenation GAM-RU block Conv	72 × 256 × 256 72 × 256 × 256 1 × 256 × 256

3.2. Filtering-Based Difference Image Generation

This section illustrates the processing steps of the DI method. DIs are generated to find a matrix representing the distance between two images. The matrix measuring change information is also called the change discriminator. Before the generation of DIs, all SLC SAR images have been pre-processed. The DI construction method consists of the NLM and INR approach. NLM filtering is used to reduce the influence of speckle noise, and its smoothing parameter can be well controlled to modulate the strength of the filtering operation. The log ratio of couples of SAR intensity patches is calculated to generate the DIs.

3.2.1. Non-Local Mean Filtering

Neighborhood information of a pixel is usually extracted from a small patch, like the common filtering methods. NLM algorithm, proposed by Buades, Coll and Morel [12], uses a search window to fuse more pixel information. Every pixel in the image is calculated by a weighted average of local neighborhood pixels. The weights characterize the similarity between neighboring and reference patches.

$$\mathrm{NLM}(x) = \frac{1}{\sum_{i\in\Omega_x} w(i,x)} \sum_{i\in\Omega_x} w(i,x) I(i) \tag{1}$$

where Ω_x represents the neighborhood of the search window of the pixel at position x; $\sum_{i\in\Omega_x} w(i,x)$ is a normalizing parameter; $w(i,x)$ is the similarity weights between patches i and x; $I(i)$ is the intensity value of the pixel at position i;

$w(i,x)$ is calculated by

$$w(i,x) = \exp\left(-\frac{\|V(i) - V(x)\|^2_{2,\alpha}}{h^2}\right) \tag{2}$$

where $\|V(i) - V(x)\|^2_{2,\alpha}$ is the Gauss weighted Euclidean distance; $V(i)$ and $V(x)$ represent the vectors of the local window at position i and x, respectively, and α is the standard deviation of the Gaussian kernel; h is a smoothing parameter and controls the decay of the weights. The larger h, the more obvious noise suppression.

3.2.2. Improved Neighborhood-Based Ratio

This study uses an INR approach to generate DIs. INR restructures the NR operator to exploit the heterogeneous information of local windows and is expected to reduce the influence of speckle noise better. The form of the INR operator is shown in the following equation.

$$\mathrm{INR}(x) = \log\left(\frac{\theta_1 \times I_1(x) + (1-\theta_1) \times u_1(x) + C}{\theta_2 \times I_2(x) + (1-\theta_2) \times u_2(x) + C}\right) \tag{3}$$

where $\mathrm{INR}(x)$ is the difference value of the pixel at position x on two images before and after changes; $I_1(x)$ represents the pixel's intensity value at position x on the image I_1; $I_2(x)$ represents the pixel's intensity value at position x on the image I_2; $u_1(x)$ is the mean of the local window at position x on the image I_1; $u_2(x)$ is the mean of the local window at position x on the image I_2. θ, shown as Equation (4), represents the heterogeneity measurement of the neighborhood area Ω_x. C is the constant for overcoming the instability, while the denominator is close to zero.

$$\theta = \frac{\sigma(x)}{\beta(x)} \tag{4}$$

where $\sigma(x)$ is the standard deviation of Ω_x; $\beta(x)$ is the mean of Ω_x.

3.3. GAM-RU Block

Compared with some attention mechanisms, i.e., CA [36] and CBAM [37], the proposed GAM-RU block incorporates the residual connection by connecting the input and output features, better solving the problems of vanishing and exploding gradients in the deep network. The output features are obtained by applying a global attention mechanism to the input features. As shown in Figure 3, The GAM-RU block consists of three parts, including spatial attention, channel attention, feature weighting, and residual connection. In the first part, the input features $F \in \mathbb{R}^{C\times W\times H}$ pass through one 3 × 3 convolutional layer for feature representations, then are processed through max pooling and average pooling along the channel axis. The output features ${F^s}_{max}$ and ${F^s}_{avg}$ are concatenated to generate efficient feature maps. Then, one 3 × 3 convolutional layer and a sigmoid function are used to obtain the final spatial weight matrix G_s, as formulated in Equation (5). The processing steps in the second part are the same as those in the first part, except for average-pooling and max-pooling operations applied along the spatial dimension, as shown in Equation (6). In the third part, the input features F are assigned the spatial attention weights and channel attention weights, simultaneously, and then a skip connection is employed to connect the output features and the input features F. The overall process can be summarized as follows:

$$G_s = \sigma\left(f^{3\times3}([maxpool(F); avgpool(F)])\right) = \sigma\left(f^{3\times3}\left([{F^s}_{max}; {F^s}_{avg}]\right)\right) \tag{5}$$

$$G_c = \sigma\left(f^{3\times3}([maxpool(F); avgpool(F)])\right) = \sigma\left(f^{3\times3}\left(\left[F^c{}_{max}; F^c{}_{avg}\right]\right)\right) \quad (6)$$

$$F_{out} = F \times G_s \times G_c + F \quad (7)$$

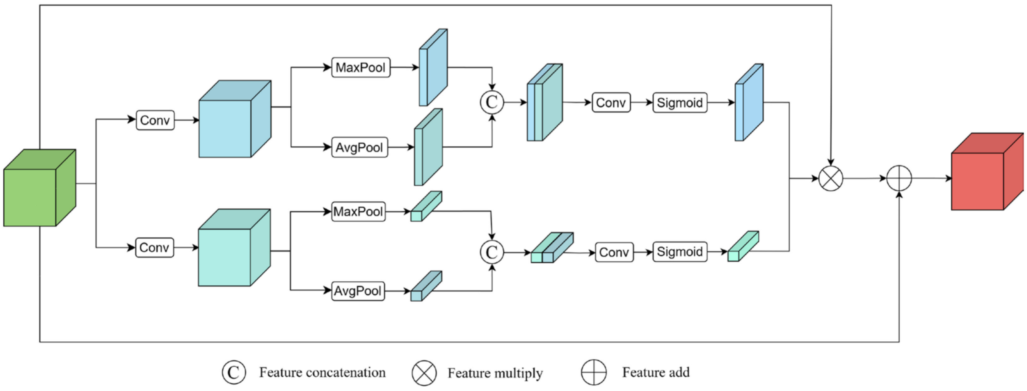

Figure 3. The structure of the GAM-RU block.

3.4. ACLA Block

To preserve detailed information on changed areas, ACLA blocks are included to assign accurate labels to each pixel through the attention mechanism. As shown in Figure 4, the ACLA block is a dual-branch structure. Two convolutional blocks are applied in the first branch for deep feature extraction. In the second branch, one convolutional block is used as the shallow feature extractor, and then the swin transformer block is used to provide better attention learning. Finally, the outputs of two branches are added to achieve accurate detection by enriching the diversity of features.

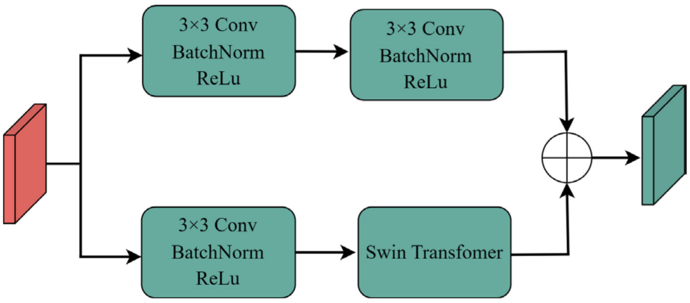

Figure 4. The structure of the ACLA block.

The core of ACLA blocks is to use a swin transformer. The transformer was first proposed for machine translation tasks and has achieved desirable results [38], but its global self-attention computation is generally unaffordable for the normal image size. Therefore, Liu et al. [34] proposed a swin transformer based on the shifted windows strategy, which improves computation efficiency by limiting global self-attention to non-overlapping local windows while allowing for crossing boundaries of the previous windows. The design element of the swin transformer is to divide the image into several patches, and each patch performs the transformer operation. For a local patch feature $X \in \mathbb{R}^{H\times W\times C}$, query Q, key K and value V are calculated by three different learnable weight matrices $W \in \mathbb{R}^{C\times C}$ shared across different patches, respectively. The self-attention operation trains feature by weighting V using the attention scores. The attention mechanism is defined as:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}} + B\right)V \quad (8)$$

where d_k is the dimension of keys, and B is the learnable relative positional encoding. The multi-head self-attention (MSA) is obtained by performing the attention function for h times and concatenating the results, in which h is set as 3 in this study (reference to most articles). As illustrated in Figure 5, the first swin transformer block consists of a window-based multi-head self-attention module (W-MSA), followed by two fully connected layers with the GELU non-linearity module. The LayerNorm (LN) layer is added before both MSA and MLP, and the residual connection is used for both modules. The next swin transformer block applies shifted window-based multi-head self-attention (SW-MSA) for better information interaction.

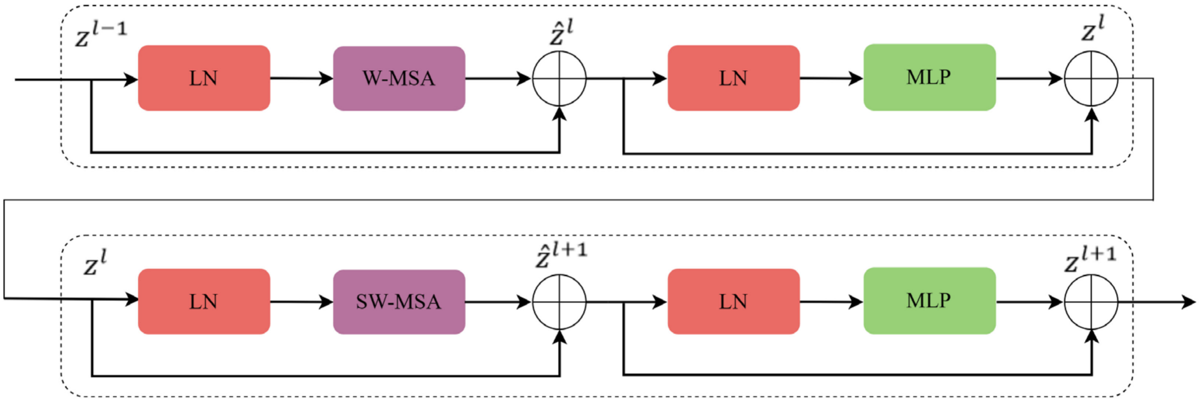

Figure 5. The illustration of connections of two successive swin transformer blocks.

In this work, we employ two successive swin transformer blocks for forward propagation, both regular and shifted windows used to enable cross-window connections. The learning process of swin transformer blocks can be expressed as:

$$\hat{z}^l = \mathrm{W-MSA}\left(\mathrm{LN}\left(z^{l-1}\right)\right) + z^{l-1} \tag{9}$$

$$z^l = \mathrm{MLP}\left(\mathrm{LN}\left(\hat{z}^l\right)\right) + \hat{z}^l \tag{10}$$

$$\hat{z}^{l+1} = \mathrm{W-MSA}\left(\mathrm{LN}\left(z^l\right)\right) + z^l \tag{11}$$

$$z^{l+1} = \mathrm{MLP}\left(\mathrm{LN}\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1} \tag{12}$$

where $\hat{z}^l$ represents the output of the W-MSA; z^l represents the output of the MLP; l denotes the l_{th} swin transformer block.

3.5. ASPP Block

The ASPP block employs multiple atrous convolutions with different dilation rates in parallel and resamples contextual information at different scales, which effectively suppresses the speckle noise of SAR images. The structure of the ASPP block is shown in Figure 6. Notably, atrous convolutions can expand the receptive field and maintain dimensional size.

In this work, ASPP performs four 3 × 3 convolutions with a dilation rate set (2, 4, 6, 8) on the previous feature maps. The BatchNorm2d and Relu operations are added after the convolution operation. Then, the fusion features obtained by channel-wise addition are subjected to one convolutional block to fuse multi-scale contextual information further.

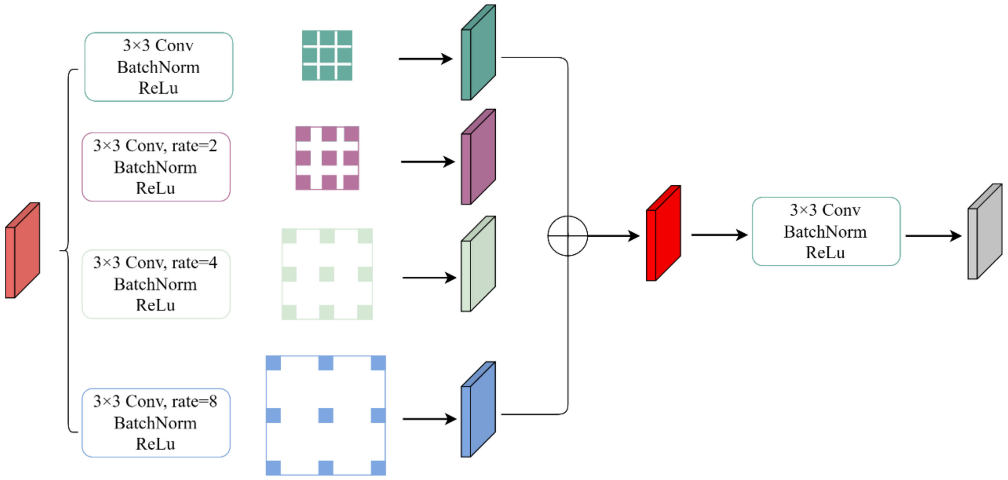

Figure 6. The structure of the ASPP block.

3.6. Implementation Details

The training environment of the proposed ADF-Net model implemented in Python using PyTorch framework is Windows 10 with NVIDIA RTX A5500 GPU and 24 GB memory. The model is trained for 100 epochs with a batch size of 8, and the learning rate of the Adam optimizer is set to 0.001 and multiplied by 80% after ten epochs. We adopted the combined loss function of binary cross-entropy L_{BCE} and Dice L_{Dice} to guide the network to learn from complex scenes and obtain reliable estimation results. The joint loss is expressed as:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i\log(p_i) + (1 - y_i)\log(1 - p_i)] \tag{13}$$

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N} y_i p_i}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} p_i} \tag{14}$$

$$L = L_{BCE} + L_{dice} \tag{15}$$

where N is the number of pixels, p_i represents the predicted category probability distribution of pixel at position i, and y_i represents the true category probability distribution of pixel at position i.

3.7. Performance Assessment

The performance evaluation of the proposed ADF-Net is critical to validate its effectiveness, as both visual interpretation and quantitative analyses are made in this study. In the visual interpretation, pre-change Sentinel-2 and post-change GaoFen-7 images are compared with generated binary change maps. For quantitative analyses, five evaluation metrics are utilized for the network model: accuracy, precision, recall, F1, and Kappa. In BCD tasks, the higher precision and recall lead to fewer false detections and omission of the changed pixels, respectively. The F1 offers a balanced assessment between precision and recall, representing the comprehensive effect of noise removal and detail preservation. The Q is an intermediate value in the calculation process, introduced to facilitate the computation of Kappa. Kappa measures the agreement between the predicted and actual classifications. Accuracy measures the proportion of correctly classified pixels among all pixels. These metrics, based on four items of true positive (TP), true negative (TN), false positive (FP), and false negative (FN), are calculated as follows:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{16}$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{17}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{18}$$

$$F1 = \frac{2 * Precision * Recall}{Precision * Recall} \tag{19}$$

$$\mathrm{Q} = \frac{(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP}) + (\mathrm{TN} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{20}$$

$$Kappa = \frac{\mathrm{TP} + \mathrm{TN} - \mathrm{Q}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN} - \mathrm{Q}} \tag{21}$$

where TP represents the number of positive samples classified to the correct classification, TN represents the number of negative samples classified to the correct classification, FP represents the number of negative samples classified to the wrong classification, and FN represents the number of positive samples classified to the wrong classification.

4. Urban BCD Performance of ADF-Net

4.1. Results of the Proposed DI Generation

Figure 7 presents the results of DI generation. Figure 7a,b are derived by performing the mean operation on pre-change and post-change TSX image time series. Due to the presence of speckle noise in image t1 and image t2, the NLM operation is performed to reduce the influence of speckle noise, as shown in Figure 7d,e. The h parameter of NLM is set as 25, based on multiple attempts for better trade-offs between denoising and retention of detailed information. Then, we utilize the INR method to generate the DIs, as shown in Figure 7f, and the C parameter of INR is set as 10. The changed areas and unchanged backgrounds can be easily distinguished from the DIs (see the red ellipses). Thus, the DIs are important features that can be used as inputs of the ADF-Net for precisely locating the changes in urban buildings.

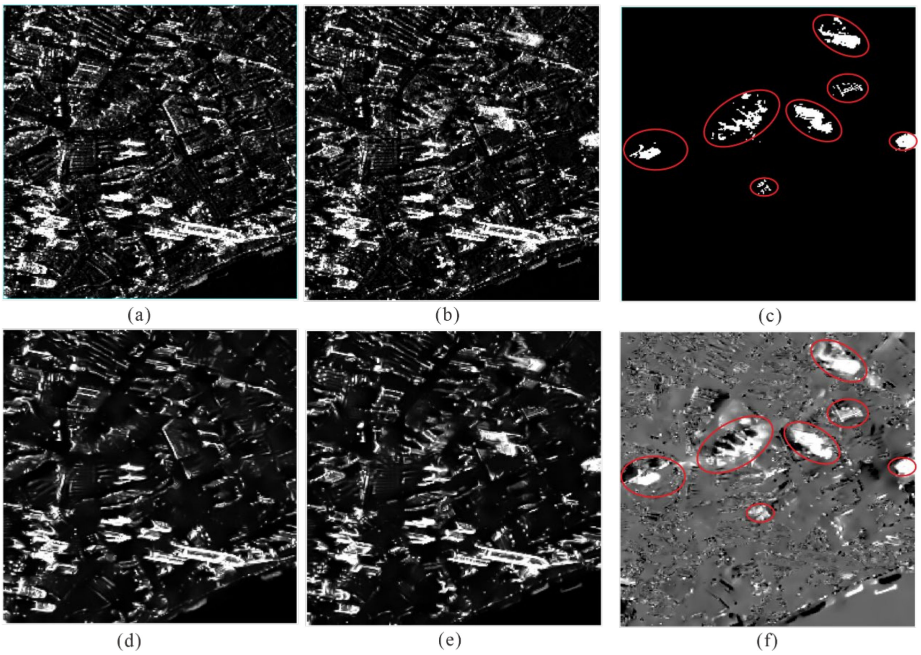

Figure 7. The illustration of DI generation. (**a**) Image t1; (**b**) image t2; (**c**) ground truth; (**d**) de-noising image t1; (**e**) de-nosing image t2; (**f**) DI.

4.2. Performance Assessment of ADF-Net

To evaluate the performance of ADF-Net on BCD, experiments are conducted on test datasets that focus on the changes in small-dense buildings. The performance of ADF-Net is compared with several state-of-the-art (SOTA) methods, including the U-Net [26], FC-EF [23], SNUNet-CD [39], A2Net [40], DMINet [41], USFFCNet [42], EATDer [43], and DRPNet [44], using evaluation metrics accuracy. The experimental settings are the same to assess the performance to eliminate the potential differences.

(1) Quantitative results

To quantitatively analyze the results, Table 3 lists the evaluation values of different methods, where the best values are in bold. The proposed ADF-Net achieves the best performance in precision (92.78%), recall (93.88%), F1 (93.10%), Kappa (93.02%), and accuracy (99.84%) of all the methods. USFFCNet shows the second-ranked performance in all methods, with F1 reaching 86.30%. Compared with USFFCNet, HSPA-Net yields an improvement of 5.29% and 7.69% for precision and recall. It denotes that ADF-Net achieves better trade-offs between denoising and retention capacity of the details. In addition, the improvements achieved by ADF-Net in F1 are 6.96% (over U-Net), 6.38% (over FC-EF), 20.77% (over SNUNet-CD), 44.27% (over A2Net), 16.14% (over DMINet), 6.80% (over USFFCNet), 14.18% (over EATDer), and 12.94 (over DRPNet). The observations demonstrate the effectiveness of ADF-Net. Two main reasons for the positive results are concluded. First, both transfomer-based and CNN-based attention mechanisms are adopted in ADF-Net, which is beneficial for extracting the local and global features of high-resolution SAR images, and the GAM-RU with a residual block enables more effective training of deeper architectures. Second, the auxiliary branch can fully utilize the information of DIs, enhancing the network's performance on BCD tasks.

(2) Visual results

The visualization comparisons of different methods are illustrated in Figure 8. ADF-Net exhibits the best visual performance, with fewer missing/false detections. Some missing detections exist in the results of the compared methods (see the green boxes in Figure 8), leading to some difficulties in detecting detailed changes, whereas the proposed ADF-Net obtains the best visual performance and shows high consistency with ground truths. In addition, many pixels are detected incorrectly in the results of U-Net, FC-EF, DMINet, USFFCNet, EATDer, and DRPNet (see the red boxes in Figure 8). At the same time, the proposed model can effectively handle the pseudo changes resulting from speckle noise. Overall, ADF-Net obtains more fine-grained detections and less salt-and-pepper noise.

(3) Model Complexity

The complexity of a model can be evaluated by calculating the number of network parameters and the FLOPs. These values for the proposed ADF-Net and the benchmark methods are presented in Table 3. The ADF-Net requires the third largest number of parameters (7.02 M), which is about the same as those of U-Net and EATDer. This is due to the dual-branch structure, which results in many parameters being trained. In addition, the FLOPs represent the time complexity. The cost of computation of ADF-Net is the second largest among all models. This is because the transformer-based attention block is introduced into our model, and it needs more time to learn to assign attention weights to different positions in the input sequence. Fortunately, owing to the unique structure, ADF-Net achieves the best BCD performance. This comparison shows that the ADF-Net improves detection accuracy by sacrificing an acceptable range of time efficiency.

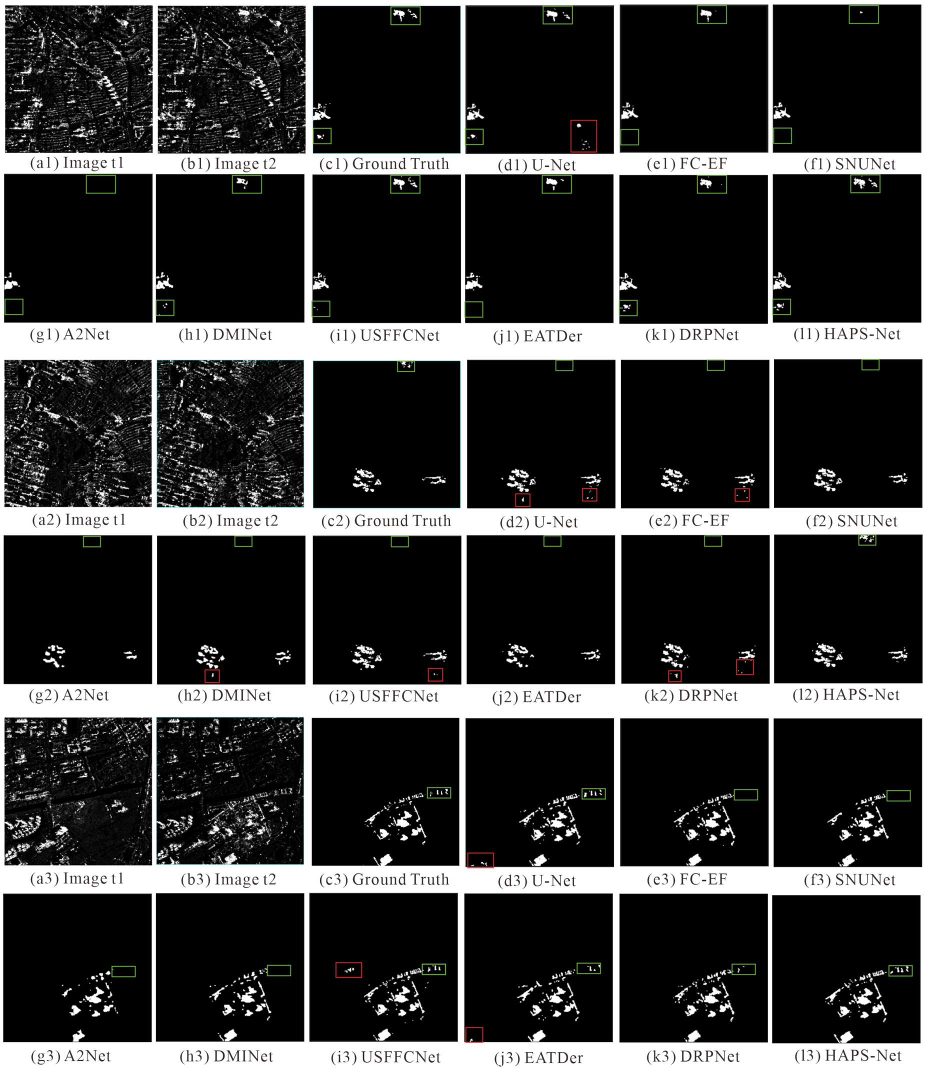

Figure 8. Visualization of urban BCD maps derived with test dataset.

Table 3. Comparison of urban BCD results obtained by different methods.

Model	Precision	Recall	F1	Kappa	Accuracy	Params (M)	FLOPs (G)
U-Net	0.8814	0.8489	0.8614	0.8597	0.9966	7.65	11.93
FC-EF	0.8849	0.8586	0.8672	0.8656	0.9967	**1.05**	**1.88**
SUNNet	0.8986	0.6426	0.7233	0.7208	0.9941	11.47	51.05
A2Net	0.8212	0.3816	0.4883	0.4849	0.9905	3.60	2.83
DMINet	0.8274	0.7311	0.7696	0.7669	0.9946	5.95	13.48
USFFCNet	0.8749	0.8619	0.8630	0.8614	0.9969	1.44	4.52
EATDer	0.8310	0.7642	0.7892	0.7866	0.9948	6.29	21.84
DRPNet	0.8220	0.7904	0.8016	0.7991	0.9950	1.61	25.24
ADF-Net	**0.9278**	**0.9388**	**0.9310**	**0.9302**	**0.9984**	7.02	45.37

4.3. Ablation Study

This section studies each part of ADF-Net from the following four aspects. First, the effectiveness of ACLA is discussed. Then, the contributions of the proposed GAM-RU block will be studied, and it will be compared with other blocks, such as CA and CBAM. Next, we investigate if the auxiliary branch helps BCD or not. Finally, the different loss functions are compared for the best detection performance. In addition, experimental settings and parameters are set the same according to the contents mentioned in Section 3.6.

(1) Effectiveness of ACLA

The ACLA blocks aim to allow the model to maintain global awareness while focusing more on local regions by introducing the swin transformer. It can preserve the details and efficiently filter the speckle noise of SAR images. To verify its effectiveness, the experiment results are reported in Table 4 (#1 and #2) and demonstrate that the addition of ACLA blocks increases by 0.31% in precision, 0.74% in recall, and 0.52% in F1. Figure 9a illustrates the increasing F1 against the epochs. We can observe that ADF-Net enjoys a more stable curve on the validation set than when it does not use ALCA blocks. In addition, the mean and standard deviation (SD) of F1 are obtained for batches 10–100. ADF-Net apparently gains a higher mean and lower SD of F1, demonstrating a better generalization performance.

(2) Effectiveness of GAM-RU

The GAM-RU blocks are designed to implement the global attention mechanism for features, including channel and spatial parts. To mitigate the vanishing gradient problem, the GAM-RU blocks involve a shortcut connection that allows the input to bypass more layers and be directly added to the output. Specifically, with the help of the GAM-RU blocks, our model achieves some improvement compared with those of the absence of GAM-RU, with a gain of 0.18%, 0.34%, and 0.32% for precision, recall, and F1 (#1 and # 3 in Table 4). Moreover, the GAM-RU blocks have better detection performance than the other attention mechanisms. Figure 9b shows that introducing GAM-RU blocks achieves faster convergence and a more stable curve when training epochs increase, which obtained the highest mean and lowest SD in all comparisons, suggesting the great learning and generalization ability of GAM-RU blocks.

(3) Effectiveness of Auxiliary Branch

The auxiliary branch assists in the effective propagation of gradients by introducing the DIs. DIs represent weight features, assigning higher weights to regions with significant changes and suppressing unchanged areas. As shown in Table 4 (#1 and # 6), adding an auxiliary branch significantly improves the precision by 4.64%, the recall by 6.58%, and the F1 by 5.73%, respectively, further enhancing the feature extraction capabilities. In addition, our model exhibits a more stable curve than that of not using an auxiliary branch, as illustrated in Figure 9c.

(4) Discussion on the Loss Function

In this study, we employ a hybrid loss, which combines the BCE loss and Dice loss for ADF-Net training. Comparative experiments are conducted to explore the influence of different loss functions. As seen in Table 4, ADF-Net with BCE loss performs better than that with Dice loss, but The SD of F1 with Dice loss is lower than that with BCE loss, indicating the stability of the network. Therefore, we adopt the hybrid of the Dice loss and BCE loss as the model training loss. It exhibits the best detection performance based on the stronger guiding capability of parameter updates.

Table 4. Comparison of ablation experiments on BCD results.

No.	Variants	Precision	Recall	F1	Kappa	Accuracy	F1 Mean	F1 SD
#1	ADF-Net	**0.9278**	**0.9388**	**0.9310**	**0.9302**	**0.9984**	**0.9207**	**0.0057**
				(a) Attention-Guided Cross-Layer Addition (ACLA)				
#2	w/o ACLA	0.9247	0.9314	0.9258	0.9249	0.9983	0.9178	0.0086
				(b) Global Attention Mechanism with Residual Unit (GAM-RU)				
#3	w/o GAM-RU	0.9260	0.9354	0.9278	0.9270	0.9983	0.9179	0.0098
#4	w/CA	0.9274	0.9301	0.9256	0.9247	0.9982	0.9148	0.0145
#5	w/CBAM	0.9266	0.9322	0.9267	0.9258	0.9983	0.9172	0.0101

Table 4. *Cont.*

No.	Variants	Precision	Recall	F1	Kappa	Accuracy	F1 Mean	F1 SD
(c) Auxiliary Branch (AB)								
#6	w/o AB	0.8814	0.8730	0.8737	0.8720	0.9969	0.8653	0.0079
(d) Loss Functions								
#7	BCE	0.9246	0.9351	0.9272	0.9264	0.9983	0.9169	0.014
#8	Dice	0.9257	0.9285	0.9240	0.9231	0.9982	0.9152	0.010

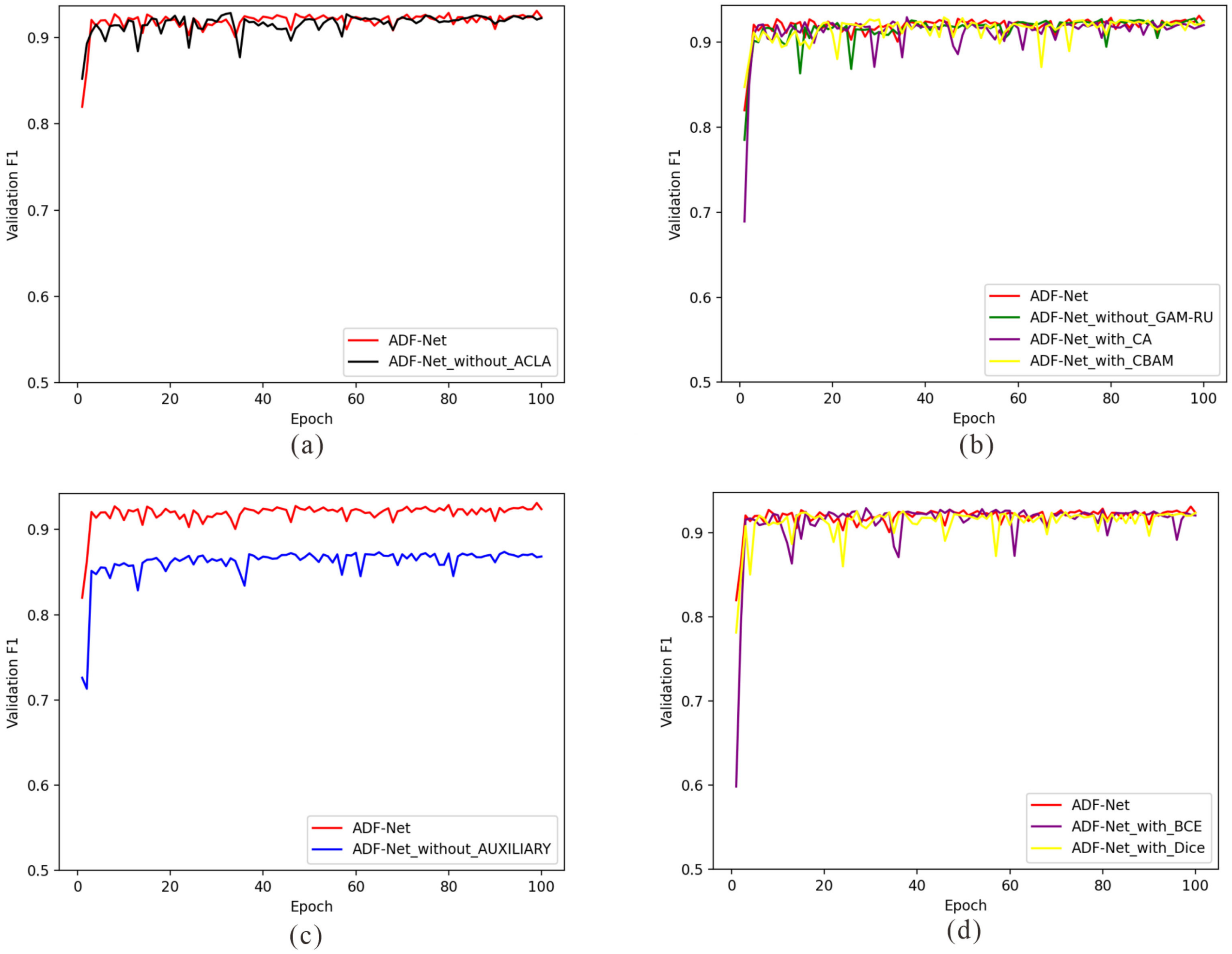

Figure 9. F1 variations against the increasing training epoch. (**a**) Validation F1 of ADF-Net with and without ACLA. (**b**) Validation F1 of ADF-Net with different CNN-based attention mechanisms. (**c**) Validation F1 of ADF-Net with and without auxiliary branch. (**d**) Validation F1 of ADF-Net with different loss functions.

5. BCD Results of Metro Line 8 in Shanghai Using TSX Images and ADF-Net

To investigate the applicability of the proposed ADF-Net method, verified by comparing different methods. The surrounding areas of Shanghai metro line 8 are selected as the study area to verify the proposed ADF-Net, because it is located in a place with many high-rise buildings that provides complex scenarios to detect small-dense building changes. The input data are six pre-processed TSX image patches, completely covering the area of metro line 8. In addition, the qualitative analysis uses the pre-change Sentinel-2 image and the post-change GaoFen-7 image to validate the BCD results.

This section selects a buffer of one kilometer centered on metro line 8 to perform BCD. The results show that, in general, many places are detected as changed areas from June 2019 to September 2021, and nine main changed zones (A–I) are displayed in yellow boxes (see Figure 10). Figure 11 show the enlarged nine changed zones and building changes highlighted with green boxes to better demonstrate the building changes.

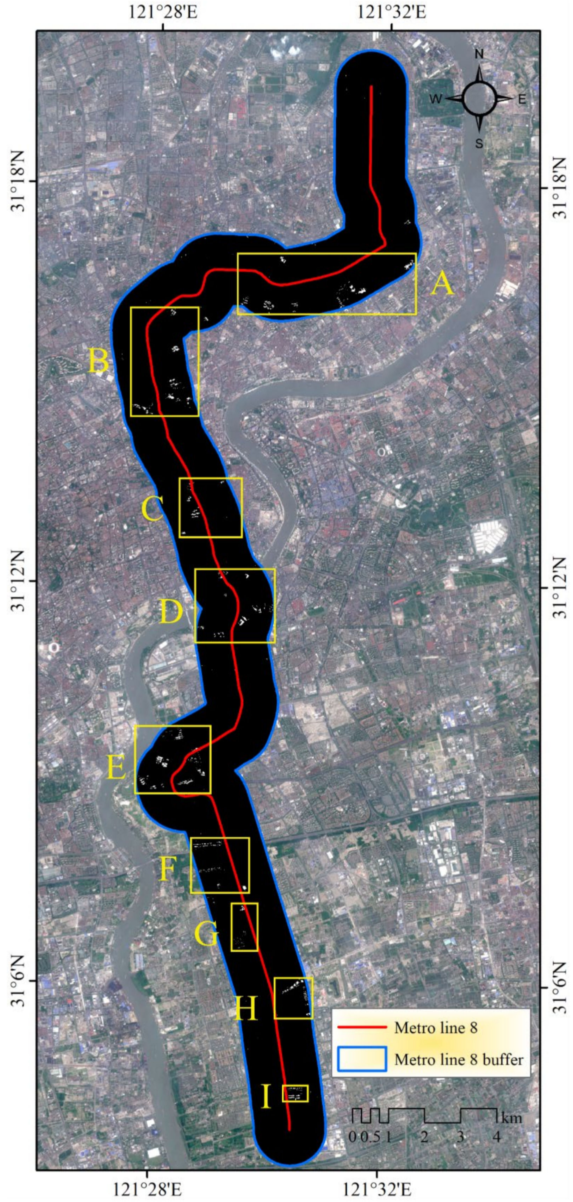

Figure 10. Urban BCD results derived by the ADF-Net model for metro line 8 in Shanghai.

The results of BCD in zones A–I imply that small-dense buildings can be successfully identified using the ADF-Net method. For SAR image BCD, reducing the influence of speckle noise is vital. The urban BCD maps with little interference from speckle noise exhibit good results. It demonstrates that the ADF-Net model can efficiently and accurately detect building changes despite complex scenarios.

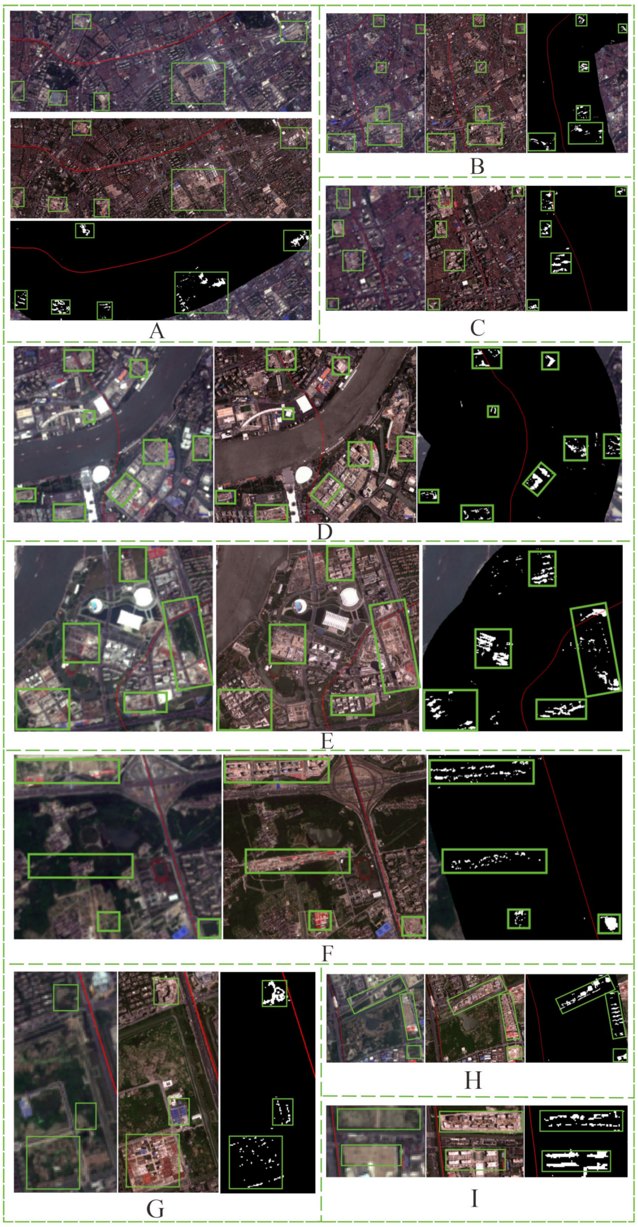

Figure 11. Urban BCD results of zones (**A**–**I**) from June 2019 to September 2021. The enlarged (**A**–**I**) zones used to better demonstrate the building changes.

6. Conclusions

This paper proposes a novel ADF-Net model using high-resolution TSX image time series for urban BCD. ADF-Net can efficiently suppress the speckle noise and enhance the feature representations of changed areas by combining the main branch and auxiliary branch networks. In the main branch, DHA-Net, ACLA blocks are utilized to obtain more detailed information. The transformer-based attention mechanism is superior in giving high weight to important pixels and suppressing the background pixels. The GAM-RU block with residual connection works well in the extraction of the deep features and avoids gradient fading. It enables the network to selectively attend to crucial channels and spatial locations, thereby improving feature discriminability. To obtain more valuable features, the auxiliary branch DMA-Net is introduced to extract the features of DIs. The NLR and INR algorithms are applied to improve the quality of DIs by suppressing the speckle noise of TSX images, and the ASPP block is utilized to realize the extraction and fusion of multi-scale features. In addition, the dual-branch structure enables ADF-Net to fuse heterogeneous information for BCD efficiently.

By applying ADF-Net to detect building changes in Shanghai, ADF-Net exhibits the best performance, generates the best building change maps, and achieves the best trade-offs between denoising and retention details compared with some SOTA methods. Ablation experiments show that introducing ACLA, GAM-RU blocks, and the auxiliary branch is essential since it can significantly improve the model performance and achieve better convergence. In addition, the BCD results of the surrounding area of metro line 8 in Shanghai demonstrate the effectiveness of the ADF-Net model and the capability to detect building changes in complex scenarios. Nine detected changed zones are consistent with the changes derived from high-resolution optical remote sensing images. In the future, coherent information on time series TSX images and spectral information on optical remote sensing images could be combined to improve BCD performance further.

Author Contributions: Conceptualization, P.C. and Q.Z.; data curation, P.C. and Q.Z.; formal analysis, P.C. and Q.Z.; funding acquisition, Q.Z.; investigation, P.C.; methodology, P.C. and Q.Z.; project administration, Q.Z.; resources, Q.Z., J.L., T.Y., X.H. and J.W.; software, P.C.; supervision, Q.Z.; validation, Q.Z.; visualization, P.C.; writing—original draft, P.C.; writing—review and editing, Q.Z. and L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Shanghai Science and Technology Development Foundation [grant number 20dz1201200], special project on Asian cooperation (the comparative study of geoenvironment and geohazards in the Yangtze River Delta and the Red River Delta), ECNU High-end Foreign Expert Project (G2022137009L), the Natural Science Foundation of China (41801337), the Research Grants of Science and Technology Commission of Shanghai Municipality through Project 18ZR1410800, and the Dragon 5 ESA-MOST GREENISH project (ID 58351). The Sentinel-2 images were downloaded from the ESA Sentinel Hub.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author, upon reasonable request.

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

References

1. Dong, X.; Mei, L.; Yang, S.; He, L.; Giorgio, I. Deformation Response Research of the Existing Subway Tunnel Impacted by Adjacent Foundation Pit Excavation. *Adv. Mater. Sci. Eng.* **2021**, *2021*, 5121084. [CrossRef]
2. Liang, R.; Xia, T.; Huang, M.; Lin, C. Simplified analytical method for evaluating the effects of adjacent excavation on shield tunnel considering the shearing effect. *Comput. Geotech.* **2017**, *81*, 167–187. [CrossRef]
3. Sun, H.; Chen, Y.; Zhang, J.; Kuang, T. Analytical investigation of tunnel deformation caused by circular foundation pit excavation. *Comput. Geotech.* **2019**, *106*, 193–198. [CrossRef]
4. Ye, S.; Zhao, Z.; Wang, D. Deformation analysis and safety assessment of existing metro tunnels affected by excavation of a foundation pit. *Undergr. Space* **2021**, *6*, 421–431. [CrossRef]
5. Gong, S.L.; Li, C.; Yang, S.L. The microscopic characteristics of Shanghai soft clay and its effect on soil body deformation and land subsidence. *Environ. Geol.* **2008**, *56*, 1051–1056. [CrossRef]
6. Shafique, A.; Cao, G.; Khan, Z.; Asad, M.; Aslam, M. Deep Learning-Based Change Detection in Remote Sensing Images: A Review. *Remote Sens.* **2022**, *14*, 871. [CrossRef]
7. Jiang, H.; Hu, X.; Li, K.; Zhang, J.; Gong, J.; Zhang, M. PGA-SiamNet: Pyramid Feature-Based Attention-Guided Siamese Network for Remote Sensing Orthoimagery Building Change Detection. *Remote Sens.* **2020**, *12*, 484. [CrossRef]
8. Peng, D.; Bruzzone, L.; Zhang, Y.; Guan, H.; Ding, H.; Huang, X. SemiCDNet: A Semisupervised Convolutional Neural Network for Change Detection in High Resolution Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5891–5906. [CrossRef]
9. Li, L.; Wang, C.; Zhang, H.; Zhang, B.; Wu, F. Urban Building Change Detection in SAR Images Using Combined Differential Image and Residual U-Net Network. *Remote Sens.* **2019**, *11*, 1091. [CrossRef]
10. Saha, S.; Bovolo, F.; Bruzzone, L. Building Change Detection in VHR SAR Images via Unsupervised Deep Transcoding. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 1917–1929. [CrossRef]
11. Zhang, X.; Liu, G.; Zhang, C.; Atkinson, P.M.; Tan, X.; Jian, X.; Zhou, X.; Li, Y. Two-Phase Object-Based Deep Learning for Multi-Temporal SAR Image Change Detection. *Remote Sens.* **2020**, *12*, 548. [CrossRef]
12. Buades, A.; Coll, B.; Morel, J.-M. Non-Local Means Denoising. *Image Process. Line* **2011**, *1*, 208–212. [CrossRef]
13. Zheng, Y.; Zhang, X.; Hou, B.; Liu, G. Using Combined Difference Image and *k*-Means Clustering for SAR Image Change Detection. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 691–695. [CrossRef]

14. Bovolo, F.; Marin, C.; Bruzzone, L. A Hierarchical Approach to Change Detection in Very High Resolution SAR Images for Surveillance Applications. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 2042–2054. [CrossRef]
15. Celik, T. A Bayesian approach to unsupervised multiscale change detection in synthetic aperture radar images. *Signal Process.* **2010**, *90*, 1471–1485. [CrossRef]
16. Inglada, J.; Mercier, G. A New Statistical Similarity Measure for Change Detection in Multitemporal SAR Images and Its Extension to Multiscale Change Analysis. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 1432–1445. [CrossRef]
17. Gong, M.; Cao, Y.; Wu, Q. A Neighborhood-Based Ratio Approach for Change Detection in SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2012**, *9*, 307–311. [CrossRef]
18. Gong, M.; Zhou, Z.; Ma, J. Change detection in synthetic aperture radar images based on image fusion and fuzzy clustering. *IEEE Trans. Image Process.* **2012**, *21*, 2141–2151. [CrossRef] [PubMed]
19. Krinidis, S.; Chatzis, V. A robust fuzzy local information C-Means clustering algorithm. *IEEE Trans. Image Process.* **2010**, *19*, 1328–1337. [CrossRef] [PubMed]
20. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
21. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [CrossRef]
22. Chakraborty, D.; Singh, S.; Dutta, D. Segmentation and classification of high spatial resolution images based on Hölder exponents and variance. *Geo-Spat. Inf. Sci.* **2017**, *20*, 39–45. [CrossRef]
23. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440. [CrossRef]
24. Li, S.; Zhao, X.; Zhou, G. Automatic pixel-level multiple damage detection of concrete structure using fully convolutional network. *Comput.-Aided Civ. Infrastruct. Eng.* **2019**, *34*, 616–634. [CrossRef]
25. Hao, C.; Shi, Z. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sens.* **2020**, *12*, 1662. [CrossRef]
26. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
27. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, 20 September 2018*; Springer International Publishing: Cham, Switzerland, 2018; Volume 11045, pp. 3–11. [CrossRef]
28. Ding, Q.; Shao, Z.; Huang, X.; Altan, O. DSA-Net: A novel deeply supervised attention-guided network for building change detection in high-resolution remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *105*, 102591. [CrossRef]
29. Yang, L.; Chen, Y.; Song, S.; Li, F.; Huang, G. Deep Siamese Networks Based Change Detection with Remote Sensing Images. *Remote Sens.* **2021**, *13*, 3394. [CrossRef]
30. Basavaraju, K.S.; Sravya, N.; Lal, S.; Nalini, J.; Reddy, C.S.; Dell'Acqua, F. UCDNet: A Deep Learning Model for Urban Change Detection From Bi-Temporal Multispectral Sentinel-2 Satellite Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5408110. [CrossRef]
31. Gao, Y.; Gao, F.; Dong, J.; Li, H.-C. SAR Image Change Detection Based on Multiscale Capsule Network. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 484–488. [CrossRef]
32. Chen, G.; Zhao, Y.; Wang, Y.; Yap, K.-H. SSN: Stockwell Scattering Network for SAR Image Change Detection. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 4001405. [CrossRef]
33. Wang, R.; Wang, L.; Wei, X.; Chen, J.-W.; Jiao, L. Dynamic Graph-Level Neural Network for SAR Image Change Detection. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 4501005. [CrossRef]
34. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002.
35. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef]
36. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13708–13717.
37. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; pp. 3–19.
38. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. Available online: https://arxiv.org/pdf/1706.03762.pdf (accessed on 30 June 2017).
39. 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*, 8007805. [CrossRef]
40. Li, Z.; Tang, C.; Liu, X.; Zhang, W.; Dou, J.; Wang, L.; Zomaya, A.Y. Lightweight Remote Sensing Change Detection With Progressive Feature Aggregation and Supervised Attention. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5602812. [CrossRef]

41. Feng, Y.; Jiang, J.; Xu, H.; Zheng, J. Change Detection on Remote Sensing Images Using Dual-Branch Multilevel Intertemporal Network. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4401015. [CrossRef]
42. Lei, T.; Geng, X.; Ning, H.; Lv, Z.; Gong, M.; Jin, Y.; Nandi, A.K. Ultralightweight Spatial–Spectral Feature Cooperation Network for Change Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4402114. [CrossRef]
43. Ma, J.; Duan, J.; Tang, X.; Zhang, X.; Jiao, L. EATDer: Edge-Assisted Adaptive Transformer Detector for Remote Sensing Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5602015. [CrossRef]
44. Chen, H.; Pu, F.; Yang, R.; Tang, R.; Xu, X. RDP-Net: Region Detail Preserving Network for Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5635010. [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.