

SARAS-Net: Scale and Relation Aware Siamese Network for Change Detection

Chao-Peng Chen¹, Jun-Wei Hsieh^{1*}, Ping-Yang Chen², Yi-Kuan Hsieh¹, Bor-Shiun Wang²

¹College of Artificial Intelligence and Green Energy, National Yang Ming Chiao Tung University, Taiwan.

²Department of Computer Science, National Yang Ming Chiao Tung University Taiwan.

f64051041@gmail.com, [jwhsieh, pingyang.cs08, khjhsnaughty.ai10, eddiewang.cs10]@nycu.edu.tw

Abstract

Change detection (CD) aims to find the difference between two images at different times and outputs a change map to represent whether the region has changed or not. To achieve a better result in generating the change map, many State-of-The-Art (SoTA) methods design a deep learning model that has a powerful discriminative ability. However, these methods still get lower performance because they ignore spatial information and scaling changes between objects, giving rise to blurry or wrong boundaries. In addition to these, they also neglect the interactive information of two different images. To alleviate these problems, we propose our network, the Scale and Relation-Aware Siamese Network (SARAS-Net) to deal with this issue. In this paper, three modules are proposed that include relation-aware, scale-aware, and cross-transformer to tackle the problem of scene change detection more effectively. To verify our model, we tested three public datasets, including LEVIR-CD, WHU-CD, and DSFIN, and obtained SoTA accuracy. Our code is available at <https://github.com/f64051041/SARAS-Net>.

Introduction

Change detection is a critical and challenging research topic in computer vision and remote sensing. This issue aims to find the difference between two images at different times and output a change map to represent whether the region has changed or not, as shown in Figure 1. The change detection task has been widely used in many applications, such as urban expansion (Lu, Moran, and Hetrick 2011), damage assessment (Xu et al. 2019), and land cover monitoring (Holley, Veraverbeke, and Hook 2014). To generate a change map, most traditional methods focus on detecting the changed pixels and classifying them. However, these results often come with low accuracy because of some noise, including different light intensity and surface colors. Hence, designing a good network with powerful discrimination to solve these problems is crucial.

With the development of deep learning, most existing methods have been proposed with powerful CNN models to tackle change detection. They have better performance than traditional methods because their outstanding discriminative

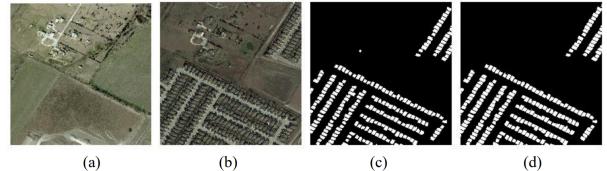


Figure 1: Result of our model for change detection in the LEVIR-CD dataset. (a) and (b) input remote sensing images, (c) ground truth, and (d) prediction result of our model.

ability can extract more useful features from images. However, these methods still face some problems when analyzing the change region. For example, FCN (Jaturapitpornchai et al. 2019) uses the U-net model to detect the region that constructs new buildings. Although it can roughly indicate the position of newly built constructions, it gets low performance because it ignores spatial information and different scale changes between objects. Although SNUNet (Fang et al. 2022) focuses on processing multi-scale features to tackle the scaling changes of objects through an ECAM (Ensemble Channel Attention Module). However, this ECAM considers only channel attention and ignores the spatial relations between pixels to generate the change maps, so many unexpected regions with seasonal changes in vegetation are also detected. To punish attention to unchanged feature pairs and increase attention to changed feature pairs, some methods (Liu et al. 2021a; Zhang et al. 2020; Peng et al. 2021) have used attention mechanisms, such as channel attention and spatial attention, to improve the detection result. However, these networks emphasize each pixel's channel importance to make the extractor more effective; it still neglects the cross-relation between features that are generated by two remote sensing images. In contrast to these networks, BIT (Chen, Qi, and Shi 2022a) uses the transformer (Vaswani et al. 2017) to encode high-level concepts of the change of interest by a set of semantic tokens and then fuses them with the original deep features to generate the expected binary change map. Though it applies attention mechanisms and considers the relationship between two features, it does not consider performing some convolution operations to fine-tune change maps after feature subtraction.

From the above discussions, we summarize the problems

*Corresponding author.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

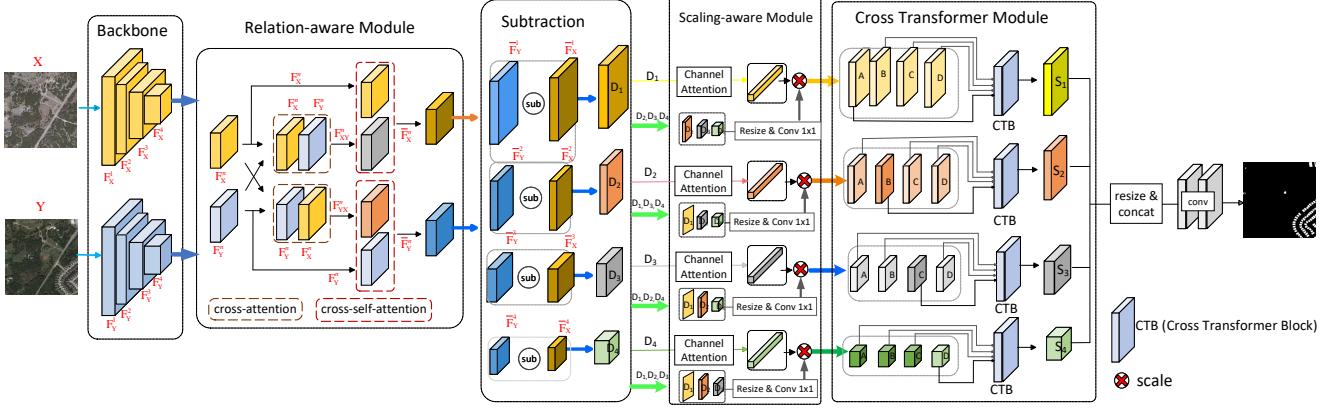


Figure 2: Overview of the proposed scale- and relation-aware siamese network.

they encountered, including multi-scale objects, the relation between two images, and focusing on important channels. In addition, we find that there is another factor that influences their effectiveness. That is, all of them perform convolution before or after the distance of features. The first type of methods, for instance, FCN and SNUNet, initially concatenates two input images and then uses some convolution operations on the concatenated map to output the change result. The second type of methods, like DASNet (Chen et al. 2021) and BIT, initially performs some convolution operations on input images and subsequently subtracts their feature maps through a few convolution layers to generate the change map. However, in our experiments, we find that performing all operations before and after feature subtraction can obtain more information and result in a better result.

This paper proposes a new network with some mechanisms to solve the disadvantages of the above methods, as shown in Figure 2. First, our network performs both operations before and after feature subtraction, respectively, using the relation-aware module before subtraction and using the scale-aware and cross-transformer modules after subtraction. The goal of the relation-aware module is to enhance the interactive relationships between feature maps extracted from two input images to improve the discrimination ability of features for change detection. Then after feature subtraction, the scale-aware attention module computes cross-scaling attention on subtraction maps to deal with the problem of scene change caused by objects with multiple sizes. Finally, the cross-transformer module, which fuses the multi-level features, aims to pay more attention to spatial information and separate foreground and background easily, thus reducing false alarms.

To solve the change detection problem and improve features' discrimination abilities, our model contributions in this paper are as follows:

- We propose a siamese network that performs both operations before and after feature subtraction on two input images to detect the change region and obtain state-of-the-art performance on the remote sensing datasets.
- We propose the relation-aware module to make the fea-

tures, which are extracted before subtraction, have more information exchanges to improve their discrimination ability for change detection.

- We propose the scale-aware module, which makes the features focus on more important channels by computing cross-scaling attention on subtraction maps, to more effectively detect changes caused by different scaled objects.
- We propose the cross-transformer module to easily separate changed pixels from unchanged ones by a self-attention mechanism.

Related Work

Change detection

In the literature, many frameworks have been proposed for change detection. According to the processed units, they can be further divided into two classes, respectively, pixel-based (Cao et al. 2014; Celik 2009; Wu et al. 2017) and object-based methods (Gil-Yepes et al. 2016; Zhang, Li, and Cui 2018). The first one generates a change map by comparing two coregistered images captured at different time pixel by pixel and then using a thresholding or clustering method to determine the locations of changed regions. However, many false detections will be produced due to many irrelevant changes such as lighting, weather, atmospheric, changes in road conditions, or seasonal changes in vegetation. To alleviate the above problems, some image preprocessing tasks should be performed, such as radiation correction, geometric correction, brightness normalization, and so on. Object-based methods (Gil-Yepes et al. 2016; Zhang, Li, and Cui 2018) typically use structural and textural features to segment raw images into objects and then obtain their change maps. Although object-based methods consider spatial information to improve change detection performance, they are sensitive to registration errors, objects' shadows, lighting conditions, and the performance of their adopted segmentation algorithms.

Using a convolutional neural network (CNN) to extract deep features for change detection has become more popular in recent years and performs much better than hand-crafted

features. Basically, CNN-based methods adopt an encoder-decoder-like (or U-Net) architecture to generate the desired change map, where the encoder converts the image pairs to various feature pyramids from which the decoder generates the final change map. For example, FCN (Jaturapitpornchai et al. 2019) uses two SAR images to detect regions that include new buildings. Furthermore, an improved UNet++ architecture (Peng, Zhang, and Guan 2019) was proposed to obtain the final binary change map by concatenating co-registered image pairs as inputs. In addition to UNet-based methods, more Siamese architectures with various attention mechanisms were proposed for change detection. For example, FC-Siam-diff (Caye Daudt, Le Saux, and Boulch 2018) uses a symmetric network to extract two temporal features and subtract them to obtain the change map. The difference map is the most intuitive feature to reveal the changes in bitemporal images, although the existence of spectral and position errors will produce many false alarms. Thus, more frameworks are proposed for change detection based on the difference in images. For example, in (Chen et al. 2021), a dual-attentive fully convolutional Siamese Networks (DAS-Net) was proposed to obtain more discriminant features by focusing both channel and spatial attention together for change detection. In (Zhang et al. 2020; Peng et al. 2021), attention maps were calculated not only from raw image pairs but also from their difference maps to assign changed pixels with higher importance but unchanged pixels with lower importance. BIT (Chen, Qi, and Shi 2022a) proposed an effective transformer-based change detection architecture and paid more attention to the changed regions.

Transformers

Transformer (Vaswani et al. 2017) is a new attention-based method for machine translation and has achieved promising performance in computer vision (Rezatofighi et al. 2019; Chi, Wei, and Hu 2020). For example, with the Vision Transformer (ViT) (Dosovitskiy et al. 2020) as the backbone, more informative features can be extracted than using spatial convolution layers networks, such as ResNet (Ramachandran et al. 2019). ViT outperforms and achieves better accuracy than CNN-based methods (Ramachandran et al. 2019; Liu et al. 2016; Bochkovskiy, Wang, and Liao 2020) in several vision tasks including object detection (Rezatofighi et al. 2019) and image segmentation (Chi, Wei, and Hu 2020). It splits the original image into non-overlapping medium-sized patches and computes their self-attentions to get more discriminant features. Although it performs well, it is very time-consuming. To alleviate this inefficiency, Swin transformer (Liu et al. 2021b) uses a smaller window size and patch interaction mechanism to achieve better speed-accuracy trade-off in image classification. This paper will employ this self-attention mechanism to strengthen feature maps not only at the same scales but also cross scales to well detect areas of changed pixels with various sizes.

Methodology

Overview

Most SoTA methods (Chen, Qi, and Shi 2022b; Chen and Shi 2020a; Chen et al. 2021) used the attention module to

enhance features before image subtraction from image pairs or fewer methods (Bai et al. 2022) enhanced the difference map after subtraction. More importantly, the above methods calculated attention from features layer by layer only at the same scales. Many miss predictions to small change areas and false alarms to large irrelevant changes will be produced. Two key ideas are proposed in this paper to alleviate the above scaling problems. The first one is to calculate attention for enhancing features from image pairs not only before subtraction but also from the difference map after subtraction. The second one is to calculate attention from deep features layer by layers not only at the same scales but also cross scales to well detect change areas even with various sizes. Our model is shown in Figure 2 and its details are shown in Algorithm 1. To compare two temporal high-resolution remote sensing images, we design a Siamese network model to extract their features. Firstly, a relation-aware model is proposed and applied to image pairs before subtraction to fuse feature maps and enhance their discriminant capabilities for change detection. After subtraction, we use the scale-aware module to calculate channel attentions on feature maps to not only at the same scale, but also other scales to deal with the scale-aware problem in change detection. After channel weighting by the scale-aware module, we use the cross-transformer to further cross-fuse features from different layers to capture more spatial and semantic information for detecting regions with change caused by objects with various sizes. Our model obtains SoTA performance on three public remote sensing datasets, including LEVIR-CD, WHU-CD, and DFIN.

Relation-aware module

Let X and Y be the two input images. To compare X and Y , a backbone such as ResNet (He et al. 2016) is first adopted to extract two feature pyramids F_X and F_Y , respectively. Let F_X^n and F_Y^n denote the feature maps of F_X and F_Y at the n th layer. To better detect changed pixels, F_X^n and F_Y^n will be improved by two mechanisms, respectively, cross-attention and cross-self-attention. Detailed operations, shown in Figure 3, are composed of sequentially connected encoder layers. The input features F_i and F_j initially produce queries Q_i and Q_j , keys K_i and K_j , and value V_i and V_j , then they are passed to the attention layer. After generating the attention weight by the dot product between the query Q_i and the key vector K_j , the attention information is retrieved by the product of the value vector V_j and the attention weight. The attention layer is denoted as:

$$A(Q_i, K_j, V_j) = \text{softmax}(Q_i K_j^T) V_j. \quad (1)$$

When the attention vector is obtained, we concatenate it and the input feature F_i to get a new feature $F_{i,j}$ as follows.

$$F_{i,j} = F_i + A(Q_i, K_i, V_i) + A(Q_i, K_j, V_j). \quad (2)$$

Similarly, we can also obtain $F_{j,i}$. In the end, the output vector is computed by a 3×3 convolution and normalization. For the cross-attention module, F_i and F_j have the same size. As to the cross-self-attention module, F_i is generated

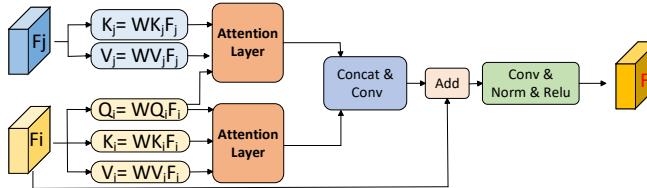


Figure 3: Detailed operations of the cross-self-attention module and cross-attention module.

by the cross-attention module, while F_j is from the original image. Both two modules can strengthen the features with more discriminant capabilities for change detection.

Scale-aware module

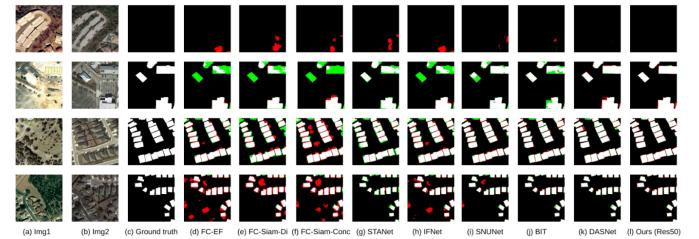
After the relation-aware module, the enhanced feature maps $\{\bar{F}_X^n\}$ and $\{\bar{F}_Y^n\}$ can be obtained from F_X^n and F_Y^n . By subtracting \bar{F}_X^n with \bar{F}_Y^n , the subtraction result D_n can be obtained, i.e. $D_n = \text{abs}(\bar{F}_X^n - \bar{F}_Y^n)$. Given the set of difference maps $\{D^i\}$, this section will propose a scale-aware module to enhance their discriminant capabilities for change detection. Different from other attention methods which only consider attention on feature maps at the same scales, this module calculates attention on feature maps not only at the same scale but also on other scales to deal with the scale-aware problem in change detection.

First, a global average pooling is applied to D_n to form a $1 \times C$ vector. It is then followed by a 1×1 convolution and activated a Sigmoid function to form a $1 \times C$ attention vector U_n . For all $\{D_n\}$, they will be resized to have the same size as D_n with a bilinear interpolation operation which is followed by a 1×1 convolution. For each resized D_m , their channels are then weighted by U_n to form a new feature map D_m^n . For the n th layer, all $\{D_m^n\}$ are then sent to the cross-transformer to generate a scale-aware feature map.

Cross transformer module

After the scale-aware module enhances each difference feature maps D_n , this section proposes a CTB(Cross-Transformer Block), as shown in Figure 7, to generate a scale-aware feature map for better scene detection. Let the inputs of CAB be D_a, D_b, D_c , and D_d which are the subtraction results from different scales. Given D_a , we train three matrices W_Q^a, W_K^a , and W_V^a to map it to the query Q_a , the key K_a , and the value V_a , respectively. Similarly, given D_b, D_c , and D_d , we can train the linear matrices $W_K^b, W_V^b, W_K^c, W_V^c, W_K^d$, and W_V^d to get $K_b, V_b, K_c, V_c, K_d, V_d$, respectively. Then, based on Q_a , we can train the cross-scale attentions β_m between it and all keys V_m for $m = a, b, c, d$, where β_m is obtained as follows:

$$\beta^m = \frac{\text{Sum}(Q_a \otimes K_m)}{\sum_{m=a,b,c,d} \text{Sum}(Q_a \otimes K_m)}. \quad (3)$$



(a) LEVIR-CD
(b) WHU-CD
(c) DSIFN-CD

Figure 4: Using visualization results to compare our model with other methods on three open datasets; that is, from top to bottom: LEVIR-CD, WHU-CD, and DSIFN-CD.

Then, all the V_m will be combined to form S_a as follows:

$$S^a = D_a + \sum_{m=a,b,c,d} \beta_m V_m, \quad (4)$$

where a skin connection is used to avoid the problem of vanish gradient by adding D_a to Eq.(5). For the n th layer, by taking $\{D_m^n\}$ as the inputs, CTB will output S_n . All $\{S_n\}_{n \neq 1}$ will be normalized to have the same size to S_1 with a bilinear interpolation operation. They are further concatenated together and followed by a 3×3 convolution to form the change classifier, $G : R^{H \times W \times 4C} \rightarrow R^{H \times W \times 2}$. With G , the final predicted change probability map P can be generated via a Softmax function as follow:

$$P = \text{Softmax}(G(S_1 \odot S_2 \odot S_3 \odot S_4)), \quad (5)$$

where \odot is a concatenation operation. Algorithm 1 shows the details of our SARAS-Net for change detection.

Network Details

To extract useful features from two input images, we use two modified backbones, ResNet18 and ResNet50 (He et al. 2016), respectively. Unlike the original ResNet18 and ResNet50, we use four feature maps that are extracted from the last four stages and replace the convolutions used the last two stages with stride 2 to 1 to achieve a speed-accuracy

Table 1: Compare our model with other methods on LEVIR-CD, WHU-CD, and DSIFN-CD dataset

Model	LEVIR-CD					WHU-CD					DSIFN-CD				
	Pre.	Rec.	F1	IoU	OA	Pre.	Rec.	F1	IoU	OA	Pre.	Rec.	F1	IoU	OA
FC-EF	84.82	77.55	81.02	68.11	97.99	77.24	68.88	72.82	57.26	97.82	61.80	57.75	59.71	42.56	86.77
FC-Siam-Di	86.73	77.52	81.87	69.31	98.11	71.61	73.40	72.49	56.86	97.64	68.44	58.27	62.95	45.93	88.35
FC-Siam-Conc	79.85	83.00	81.39	68.62	97.90	76.94	69.74	73.17	57.69	97.83	59.08	62.80	60.88	43.76	86.30
STANet	89.47	83.31	86.28	75.88	98.54	90.62	86.26	88.38	79.19	99.04	51.48	36.40	42.65	27.11	83.38
IFNet	83.12	79.58	81.31	68.51	97.98	75.92	71.53	73.66	58.31	97.83	63.75	55.36	59.26	42.11	87.08
SNUNet	89.43	87.72	88.57	79.48	98.75	91.88	84.57	88.07	78.69	99.03	64.15	57.09	60.41	43.28	87.30
BIT	89.35	89.56	89.46	80.92	98.83	89.40	90.03	89.72	81.35	99.12	56.36	62.79	59.40	42.25	85.43
DASNet	90.60	91.38	90.99	83.47	99.09	88.23	84.62	86.39	76.04	95.30	60.10	56.53	58.26	41.10	86.25
SARAS-Net (V1)	91.48	89.35	90.40	82.49	98.95	91.41	89.58	90.48	82.62	98.96	64.48	64.98	64.73	47.85	88.05
SARAS-Net (V2)	91.97	91.85	91.91	84.95	99.10	92.94	89.12	90.99	83.47	99.25	67.65	67.51	67.58	51.04	89.01

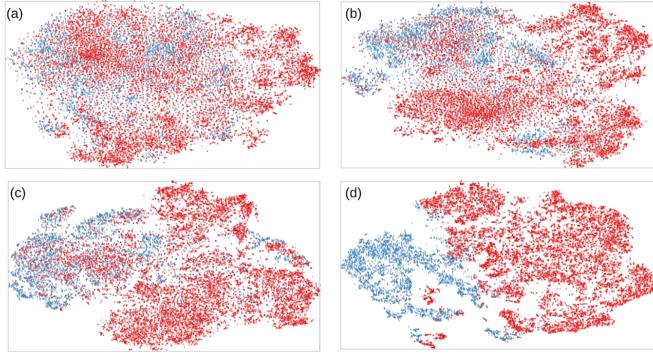


Figure 5: Results of visualization of features enhanced by different modules. (a) Subtraction result without using any module. (b) Result after adding the relation-aware module. (c) Result after adding the scale-attention module. (d) Result after using the cross-transformer module.

trade-off. In our modified ResNet50, the channel sizes for the four features are 256, 512, 1024, and 2048, respectively. To save GPU computation, we reduce the channels of four feature maps to 64/128/256/512 by 1×1 convolution. We call our model SARAS-Net (V1) and SARAS-Net (V2) when using ResNet18 and ResNet50.

Loss function

We use the cross-entropy loss function to optimize our network parameters and minimize loss value in the training stage. The loss is formulated as follows:

$$L = \frac{1}{N} \sum_{n=1}^N l(P_n, Y_n), \quad (6)$$

where Y_n is the class value, which is 0 or 1, representing whether this pixel changes or not, N is the number of pixels, P_n is the prediction value generated by our network, and $l(P_n, Y_n) = -Y_n \log(P_n) - (1 - Y_n) \log(1 - P_n)$ is the cross-entropy loss.

Experiments and Results

Three datasets, LEVIR-CD (Chen and Shi 2020b), DSIFN-CD (Zhang et al. 2020), and WHU-CD (Ji, Wei, and Lu 2019) were used to evaluate the performance of our model.

Algorithm 1: SARAS-Net for change detection

```

Input: Two temporal remote sensing images ( $X, Y$ )
Output: Change map  $M$ 
1: Step1: Feature Extraction
2:  $F_X = CNN(X); F_Y = CNN(Y);$ 
3: Step2: Relation-aware
4: for layer  $n$  do
5:    $(\bar{F}_X^n, \bar{F}_Y^n) = \text{Relation-Aware-module}(F_X^n, F_Y^n);$ 
6:    $D_n = \text{abs}(\bar{F}_X^n - \bar{F}_Y^n);$ 
7: Step3: Scale-aware adn Cross-Transformer
8: for layer  $n$  do
9:    $U_n = \text{Scale-Aware-Attention}(D_n);$ 
10:  for layer  $m (m \neq n)$  do
11:     $D_m^n = \text{Channel-wise}(U_n, \text{Resize}(D_m));$ 
12:     $S_n = \text{CTB}(D_1^n, D_2^n, D_3^n, D_4^n);$ 
13: Step4: Change Map Generation
14:    $P = \text{Softmax}(G(S_1 \odot S_2 \odot S_3 \odot S_4));$ 
14: return  $P$ 

```

Datasets

- LEVIR-CD: It contains 637 pairs of remote sensing images with the size 1024×1024 pixels. To reduce the computation and argument training data, the original image was cut into small patches that have 256×256 size. Finally, we obtained 7120/1024/2048 pairs of patches for training/validation/test datasets.
- WHU-CD: It contains only one pair of aerial images with an image size of 32507×15354 pixels. Then, this image is divided into small non-overlaped patches with the same 256×256 size. In the end, there are 6690/744/744 pairs of patches for the training / validation / test dataset.
- DSIFN-CD: Its contains changes in roads, buildings, and water. For each pair of images, their sizes are 512×512 pixels and cut into small non-overlaped patches with 256×256 sizes to obtain 14400/1360/192 pairs for training/validation/testing.

Implementation Details

In the training stage, we use three NVIDIA Tesla V100 GPU to implement our model and stochastic gradient descent (SGD) to optimize our model parameters. We set the momentum to 0.9 and the weight decay to 0.0005. Initially,

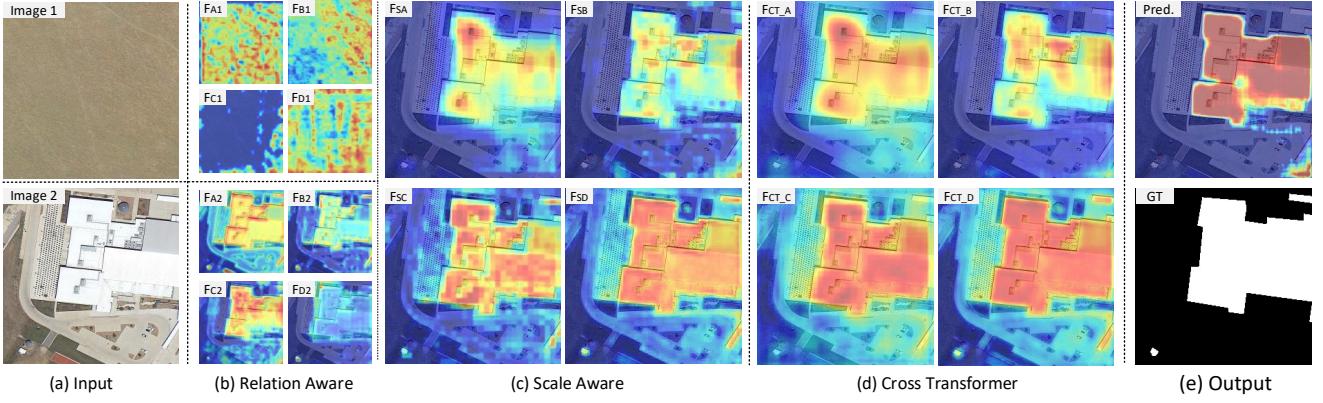


Figure 6: Example of SARAS-Net visualization by Gradcam. Red denotes higher attention values and blue denotes lower values. (a) Two input images. (b) Feature maps generated by the relation-aware module. (c) Subtraction results after adding the scale-aware module. (d) Subtraction results after using the cross-transformer module. (e) Prediction and ground truth.

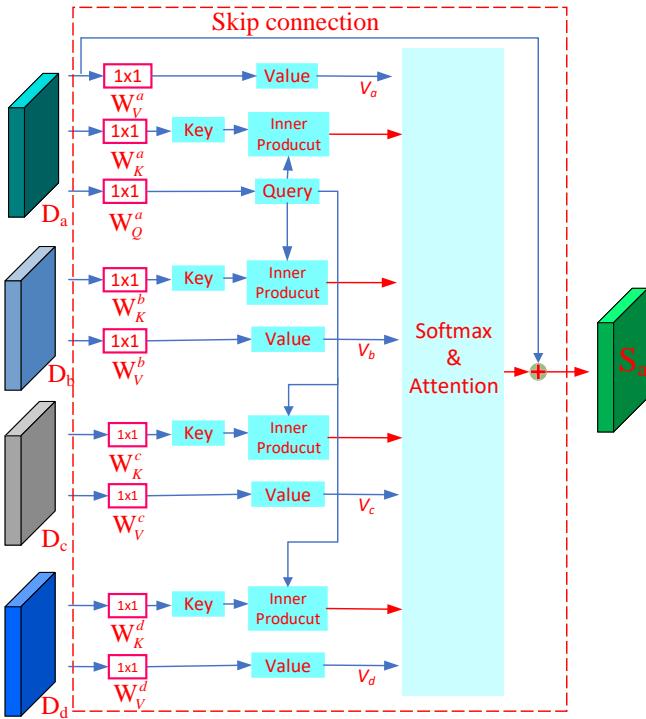


Figure 7: Detailed operations used in CTB (Cross-Transformer Block)

the learning rate is set to 0.05 and decays 0.1 times for every 50 epochs. For each epoch, we use data augmentation to obtain higher accuracy, including rescale, crop, flip, and Gaussian blur during training and use the validation dataset to choose the best training weight. Finally, we evaluate our model accuracy on the test dataset.

Experiments on dataset

We compared our model with eight SoTA methods, including FC-EF (Caye Daudt, Le Saux, and Boulch 2018), FC-

Siam-Di, FC-Siam-Conc, SNUNet, STANet (Chen and Shi 2020b), IFNet (Zhang et al. 2020), ISNet (Cheng, Wang, and Han 2022), BIT, and DASNet (Chen et al. 2021).

Table 1 illustrates performance comparisons with the SoTA methods in the three data sets, respectively. The metrics contain precision, recall, F1 score, intersection over union (IoU) of the change category, and general accuracy. Clearly, our model outperforms all SoTA methods. To visualize the prediction results, the results of different methods on the above three datasets are shown in Figure 4. Here, the white color is for true positive, the black color is for true negative, the red color is for false positive, and the green color is for false negative.

To evaluate the performance of each module, we projected high-dimensional features onto 2D maps with two colors to represent the class of pixels (see Figure 5). First, the subtraction result between two input images, shown in Figure 5(a), illustrates the changed and unchanged pixels are mixed together. Second, after the self-attention and cross-attention modules, points with the same class become closer (see Figure 5(b)). Third, as shown in Figure 5(c), the scale-attention module makes points with different classes become farther. Finally, as shown in Figure 5(d), the cross-transformer block fuses features from different layers to make changed and unchanged pixels more easily separated.

In order to better understand our model, we use gradcam (Gildenblat and contributors 2021) to visualize each module. Figure 6(a) shows the two input images. Figure 6(b) shows the feature map after adding the relation-aware module. Figure 6(c) shows the subtraction results after using the scale-aware module. From Figure 6(d), we observe that the noise in features is alleviated through the cross-transformer module. The final prediction map is shown in Figure 6(e).

Ablation study

SARAS-Net analysis. We performed an ablation study for our model on LEVIR-CD data set and changed different modules in sequence to evaluate the performance of each module. First, from Table 2, we can conclude that when we remove the relation-aware (RA) module, the number of pa-

rameters is reduced by nearly half and the FLOPs also decrease. However, the performance after removing the RA module is worse than after removing other modules. Thus, if we want a light network, we will remove the RA module. Second, we observe that the cross transformer module has higher performance since it can perfectly fuse different scale features. Figure 8 shows the changes in the loss value at each epoch during training. We observe that our model without removing any module is easier to converge.

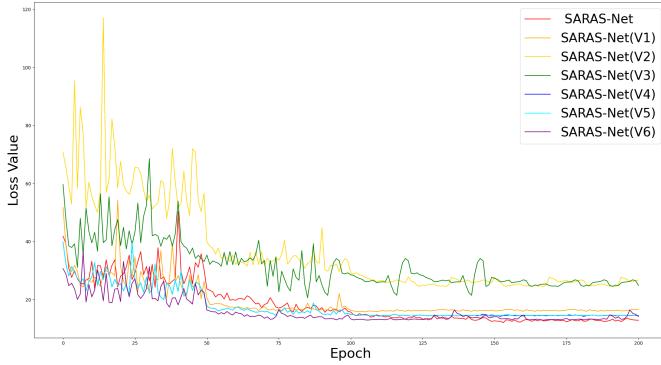


Figure 8: Ablation study of training loss. Each line represents an ablation study from Table 2.

Table 2: Ablation study of the effects when different modules are added to our model, where RA is the relation-aware module, SA is the scale-aware module, and CT is the cross-transformer module.

Model	RA	SA	CT	Param.(M)	FLOPs.(G)	F1	IoU
SARAS-Net(v1)	X	X	✓	32.33	60.64	90.63	82.86
SARAS-Net(v2)	X	✓	X	37.02	92.57	90.49	82.63
SARAS-Net(v3)	✓	X	X	47.46	76.91	90.48	82.62
SARAS-Net(v4)	X	✓	✓	42.94	111.77	91.26	83.39
SARAS-Net(v5)	✓	X	✓	52.36	108.83	91.11	83.68
SARAS-Net(v6)	✓	✓	X	53.16	128.03	90.92	83.36
SARAS-Net	✓	✓	✓	56.89	139.9	91.91	84.95

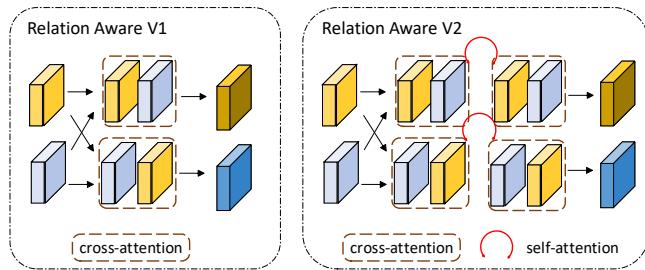


Figure 9: Ablation study of relation-aware module. Relation-aware V1 only uses cross-attention. Relation-aware V2 uses cross-attention and self-attention.

Relation-aware module analysis. To evaluate the ablation study of the relation-aware module, we performed dif-

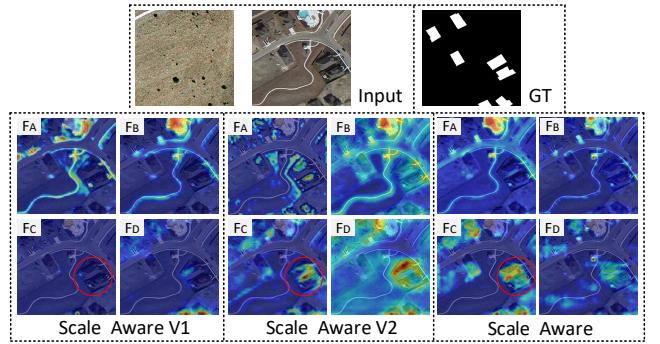


Figure 10: Ablation study of the relation-aware module. V1 uses only cross-attention and V2 uses both cross-attention and self-attention.

Table 3: Ablation study of the relation-aware module, where CA denotes cross-attention, CSA the cross-self-attention, and SA the self-attention.

Model	CA	CSA	SA	Param.(M)	FLOPs.(G)	F1	IoU
SARAS-Net	✓	✗	✗	49.22	131.77	91.54	84.41
SARAS-Net	✓	✗	✓	56.89	139.90	90.56	82.76
SARAS-Net	✓	✓	✗	56.89	139.90	91.91	84.95

ferent attention mechanisms. As shown in Figure 9, the first relation-aware version only uses cross-attention, and the second uses cross-attention initially and then replaces cross-self-attention with self-attention. From Table 3, we can observe that the first version is lighter but has a worse performance. To better understand this ablation study, we visualize the different scale features using the relation-aware module by Gradcam in Figure 10. Compared to other ablation studies of the relation-aware module, we observe that the original relation-aware module pays more attention to regions with changes. For example, as shown in Figure 10, the red circle in the original module feature F_C performs better.

Conclusion

This paper proposed a scale- and relation-aware siamese network for change detection to achieve SoTA accuracy on the LEVIR-CD, WHU-CD, and DSIFN-CD datasets. More accurately, our model obtains significant improvements in F1 scores in these datasets, respectively, 2.45, 1.27, and 4.63 points. Our method can solve the key problems of change detection encountered with most existing methods. For example, the relation-aware and scale-aware modules can resolve boundary noise generated by objects of different scales and enhance the features of interactive information. In addition, we fuse the different scale features using the cross-transformer module to get a better representation for change detection. Except for these, our main contribution is to propose a new model, which performs operations before and after feature subtraction. Through experimental evidences, our model structure has been proven to be useful in this issue.

Acknowledgments

This work was partly supported by National Science and Technology Council, Taiwan (Grant Number: MOST 109-2221-E-009-116-MY3, 110-2221-E-A49-132-MY3 and 110-2634-F-A49-006).

References

- Bai, B.; Fu, W.; Lu, T.; and Li, S. 2022. Edge-Guided Recurrent Convolutional Neural Network for Multitemporal Remote Sensing Image Building Change Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–13.
- Bochkovskiy, A.; Wang, C.-Y.; and Liao, H.-Y. M. 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv preprint arXiv:2004.10934*.
- Cao, G.; Li, Y.; Liu, Y.; and Shang, Y. 2014. Automatic change detection in high-resolution remote-sensing images by means of level set evolution and support vector machine classification. *International Journal of Remote Sensing*, 35(16): 6255–6270.
- Caye Daudt, R.; Le Saux, B.; and Boulch, A. 2018. Fully Convolutional Siamese Networks for Change Detection. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, 4063–4067.
- Celik, T. 2009. Unsupervised Change Detection in Satellite Images Using Principal Component Analysis and K-Means Clustering. *IEEE Geoscience and Remote Sensing Letters*, 6(4): 772–776.
- Chen, H.; Qi, Z.; and Shi, Z. 2022a. Remote Sensing Image Change Detection With Transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–14.
- Chen, H.; Qi, Z.; and Shi, Z. 2022b. Remote Sensing Image Change Detection With Transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–14.
- Chen, H.; and Shi, Z. 2020a. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 12(10): 1662.
- Chen, H.; and Shi, Z. 2020b. A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection. *Remote Sensing*, 12(10).
- Chen, J.; Yuan, Z.; Peng, J.; Chen, L.; Huang, H.; Zhu, J.; Liu, Y.; and Li, H. 2021. DASNet: Dual Attentive Fully Convolutional Siamese Networks for Change Detection in High-Resolution Satellite Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14: 1194–1206.
- Cheng, G.; Wang, G.; and Han, J. 2022. ISNet: Towards Improving Separability for Remote Sensing Image Change Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–11.
- Chi, C.; Wei, F.; and Hu, H. 2020. RelationNet++: Bridging Visual Representations for Object Detection via Transformer Decoder. In *Advances in Neural Information Processing Systems*, volume 33, 13564–13574. Curran Associates, Inc.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.
- Fang, S.; Li, K.; Shao, J.; and Li, Z. 2022. SNUNet-CD: A Densely Connected Siamese Network for Change Detection of VHR Images. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5.
- Gil-Yepes, J. L.; Ruiz, L. A.; Recio, J. A.; Ángel Balaguer-Beser; and Hermosilla, T. 2016. Description and validation of a new set of object-based temporal geostatistical features for land-use/land-cover change detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 121: 77–91.
- Gildenblat, J.; and contributors. 2021. PyTorch library for CAM methods. <https://github.com/jacobgil/pytorch-grad-cam>.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Hulley, G.; Veraverbeke, S.; and Hook, S. 2014. Thermal-based techniques for land cover change detection using a new dynamic MODIS multispectral emissivity product (MOD21). *Remote Sensing of Environment*, 140: 755–765.
- Jaturapitpornchai, R.; Matsuoka, M.; Kanemoto, N.; Kuzuoka, S.; Ito, R.; and Nakamura, R. 2019. Newly Built Construction Detection in SAR Images Using Deep Learning. *Remote Sensing*, 11: 1444.
- Ji, S.; Wei, S.; and Lu, M. 2019. Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1): 574–586.
- Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37. Springer.
- Liu, Y.; Pang, C.; Zhan, Z.; Zhang, X.; and Yang, X. 2021a. Building Change Detection for Remote Sensing Images Using a Dual-Task Constrained Deep Siamese Convolutional Network Model. *IEEE Geoscience and Remote Sensing Letters*, 18(5): 811–815.
- Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022.
- Lu, D.; Moran, E.; and Hetrick, S. 2011. Detection of impervious surface change with multitemporal Landsat images in an urban–rural frontier. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(3): 298–306.
- Peng, D.; Zhang, Y.; and Guan, H. 2019. End-to-End Change Detection for High Resolution Satellite Images Using Improved UNet++. *Remote Sensing*, 11(11).
- Peng, X.; Zhong, R.; Li, Z.; and Li, Q. 2021. Optical Remote Sensing Image Change Detection Based on Attention

Mechanism and Image Difference. *IEEE Transactions on Geoscience and Remote Sensing*, 59(9): 7296–7307.

Ramachandran, P.; Parmar, N.; Vaswani, A.; Bello, I.; Levskaya, A.; and Shlens, J. 2019. Stand-Alone Self-Attention in Vision Models. In *Advances in Neural Information Processing Systems*, volume 32.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30.

Wu, C.; Du, B.; Cui, X.; and Zhang, L. 2017. A post-classification change detection method based on iterative slow feature analysis and Bayesian soft fusion. *Remote Sensing of Environment*, 199: 241–255.

Xu, J. Z.; Lu, W.; Li, Z.; Khaitan, P.; and Zaytseva, V. 2019. Building Damage Detection in Satellite Imagery Using Convolutional Neural Networks.

Zhang, C.; Li, G.; and Cui, W. 2018. High-Resolution Remote Sensing Image Change Detection by Statistical-Object-Based Method. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(7): 2440–2447.

Zhang, C.; Yue, P.; Tapete, D.; Jiang, L.; Shangguan, B.; Huang, L.; and Liu, G. 2020. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166: 183–200.